



## 저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

이학박사 학위논문

A stock price process model  
reflecting dynamics of traders'  
behaviors

투자자들의 행동 역학을 반영한 주식 가격  
과정 모형

2018년 2월

서울대학교 대학원

수리과학부

김 원 세

# A stock price process model reflecting dynamics of traders' behaviors

A dissertation  
submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
to the faculty of the Graduate School of  
Seoul National University

by

Wonse Kim

Dissertation Director : Professor Hyeong In Choi

Department of Mathematical Science  
Seoul National University

February 2018

© 2018 Wonse Kim

All rights reserved.

# Abstract

## A stock price process model reflecting dynamics of traders' behaviors

Wonse Kim

Department of Mathematical Sciences  
The Graduate School  
Seoul National University

In this paper, we propose a new stock price process model that reflects the dynamics of traders' behaviors. Our model has two implications: First, in the both seller group and the buyer group, the stock price moves in favor of the minority group, not the majority group, and the smaller the minority group is, the larger the change in the price. Second, in both the seller group and the buyer group, traders follow (herd to) the behavior of the minority, and the smaller the minority group is, the larger the herding. Then, exploiting our proprietary data set, we show that our model explains the market well. We also use our model to show that we can predict stock prices via a machine-learning technique that we develop.

**Key words:** Stock price process, Informed investor, Trading skills, Machine learning, Return prediction

**Student Number:** 2013-30083

# Contents

<b>Abstract</b>	<b>i</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 The stock price process model reflecting the dynamics of traders' behavior</b>	<b>5</b>
2.1 Motivations . . . . .	5
2.1.1 Motivation based on Information theory . . . . .	5
2.1.2 Motivation based on empirical studies of traders' trading skills . . . . .	11
2.2 The model description . . . . .	12
2.2.1 SDE form . . . . .	12
2.2.2 Closed-form solution of the SDE (linear assumption) . .	14
<b>3 Empirical Results</b>	<b>17</b>
3.1 Data description . . . . .	17
3.2 Empirical results . . . . .	20
3.2.1 The dynamics of stock price processes . . . . .	20
3.2.1.1 Cardinal property of SBR . . . . .	20
3.2.1.2 Ordinal property of SBR . . . . .	21
3.2.2 The dynamics of SBR . . . . .	26
3.2.2.1 Cardinal analysis of the dynamics of traders . .	26
3.2.2.2 Ordinal analysis of the dynamics of traders . .	27
3.3 Robustness check:subperiod test . . . . .	31

## CONTENTS

<b>4</b>	<b>Return prediction via a machine learning technique</b>	<b>34</b>
4.1	Test data set description . . . . .	35
4.2	Data filtration . . . . .	35
4.3	Key predictors . . . . .	36
4.3.1	Interaction between types . . . . .	36
4.3.2	LSV herding measure of each types . . . . .	40
4.4	Other predictors . . . . .	45
4.4.1	Intraday volatility . . . . .	45
4.4.2	Predictors related to returns . . . . .	45
4.4.3	Predictors related to prices . . . . .	46
4.5	predictor model . . . . .	47
4.5.1	Model description . . . . .	48
4.5.1.1	Random forest . . . . .	48
4.5.1.2	Elastic Net . . . . .	52
4.5.1.3	Our new model: two step learning (residual fitting) . . . . .	57
4.5.2	Empirical Result . . . . .	57
4.5.2.1	SBR prediction . . . . .	58
4.5.2.2	Return prediction . . . . .	59
<b>5</b>	<b>Conclusion</b>	<b>62</b>
	<b>Abstract (in Korean)</b>	<b>68</b>

# Chapter 1

## Introduction

A geometric Brownian motion (GBM),

$$S(t) = S(0) \exp \left\{ \left( \mu - \frac{1}{2} \sigma^2 \right) t + \sigma W(t) \right\}, \quad (1.0.1)$$

is a stochastic process that satisfies a stochastic differential equation (SDE),

$$dS_t = \mu S_t dt + \sigma S_t dW_t. \quad (1.0.2)$$

GBM, one of the most popular stock price process models, is widely used in mathematical finance, particularly in the celebrated *Black-Scholes Option Pricing model*, to model stock price processes. GBM is a simple stock price process model, as it involves only two simple parameters ( $\mu$ : drift parameter,  $\sigma$ : volatility parameter) to describe stock price processes, and it does not reflect the dynamics of traders' behaviors.

However, recent empirical studies show that heterogeneous groups of investors in the market affect stock price processes in different ways: Using Finland stock market transaction data, Stoffman (2014) [34] shows that, in trades between institutions and households, stock prices move in the direction that benefits institutions, regardless of whether institutions buy from households or sell to them.



## CHAPTER 1. INTRODUCTION

Our prior studies also address the effect of traders' dynamics on stock prices. Using data from the Korea stock market and methods similar to those of Stoffman (2014) [34], Chay and Kim (2017) [10] investigate how stock prices move contemporaneously with trading activities among three types of investors: households, institutions, and foreigners. They show that, whereas prices consistently move in the direction of institutional trading in trades between individuals and institutions, prices consistently move in the direction of foreign investors at a daily horizon, no matter who is on the opposite side of their trades. Chay and Kim (2017) [10]'s results suggest that, at least in the short term, institutions predict stock prices better than individuals do, and foreign investors have an advantage in forecasting stock prices than either individuals or institutions. Chay and Kim (2017) [10] also perform the same type of analyses over a weekly horizon and show that, as in the case of their daily horizon analysis, weekly prices move to benefit institutions and foreigners when they trade with individuals. However, while Chay and Kim (2017) [10] find no significant relationship between weekly stock returns and the trading when domestic institutions trade with foreigners, they use an analysis based on the volume-weighted average price to present significant evidence that foreigners secure positive weekly returns when they trade with domestic institutions. The result of weekly analysis of Chay and Kim (2017) [10] is summarized in Figure 1.1, and Figure 1.2.

Using the same data that Chay and Kim (2017) [10] use, Chay, Kim, and Lee (2017) [11] use the measure developed by Lakonishok, Shleifer, and Vishny (1992) [27]) to measure the effect of the three types of investors' herding on the intraday volatilities of stock price processes, which is measured by the *realized variance*. They show that herding by individual investors increases intraday stock volatility, while herding by foreign institutional investors decreases intraday stock volatility. Domestic institutions' herding does not appear to affect intraday stock volatility. Chay, Kim, and Lee (2017) [11]'s empirical results, combined with information theory, suggest that, whereas individual investors tend to be noise traders in the stock market, foreign traders tend to be in-

## CHAPTER 1. INTRODUCTION

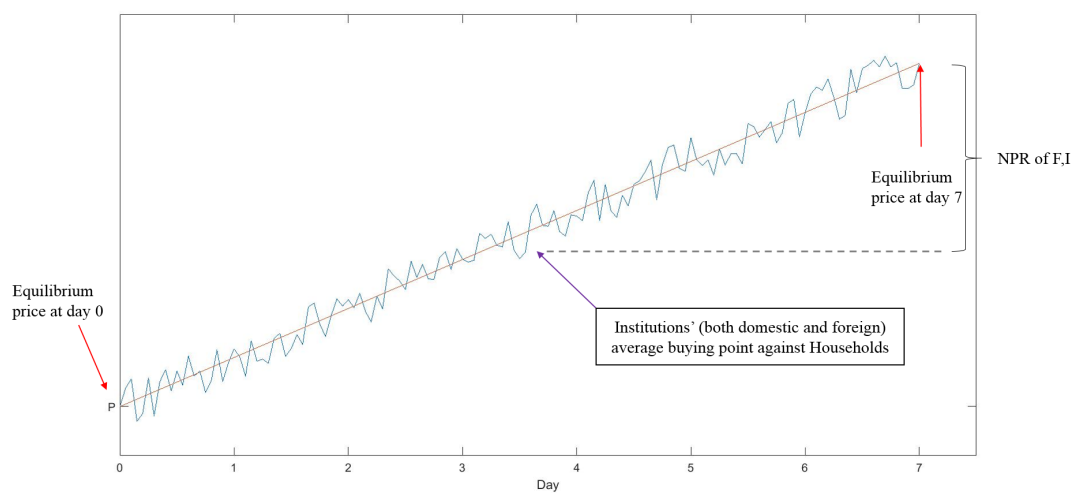


Figure 1.1

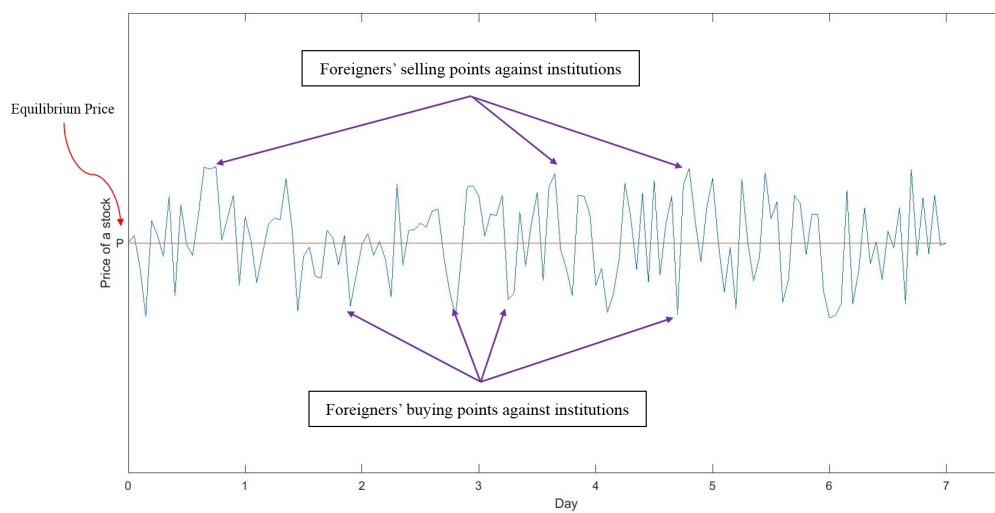


Figure 1.2

## CHAPTER 1. INTRODUCTION

formed traders in the market.

These lines of empirical studies suggest the needs for a new stock price process model that reflects the effects of the dynamics of traders' behaviors. In this paper, we suggest a stock price processes model in which stock price processes are affected by the dynamics of traders' behaviors. Then, by exploiting our proprietary data set, we show that our model reflects the financial market well. Finally, we show that, based on our model, we can predict stock returns by using a machine-learning technique that we develop.

## Chapter 2

# The stock price process model reflecting the dynamics of traders' behavior

In this chapter, we first introduce the two motivations of our new stock price processes model. We then describe the model in the form of SDE and present the closed-form solutions under some linearity assumptions.

### 2.1 Motivations

#### 2.1.1 Motivation based on Information theory

The first approach to motivate our new stock price processes model, is based on information theory. As Black (1985) [4] states that ‘Noise trading is trading on noise as if it were information’, most traders trade stocks since they regard themselves as informed traders based on their own information, whether their information is valuable or not. The value of the information (or whether the information is really information or not) becomes known by following a stock’s price movement. Under this setting of information theory, we make the following argument.

## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

**Motivation 1.** (*A motivation based on information theory*)

*The value of the information a trader group (a seller group or a buyer group) can be measured ex ante by the ratio between the sizes of the two groups.*

This argument can be supported by information theory using the Bayesian Nash equilibrium, which is based primarily on Kyle (1985) [26]. Kyle (1985) [26]'s dynamic model first shows that a single informed investor can exploit his or her monopoly power. Kyle (1985) [26] then expands the single auction model to a sequential auction model and a continuous auction model and shows that, under the some linearity conditions, the sequential auction model and the continuous auction model have unique equilibria. It is during the process of reaching the unique equilibrium that the informed trader can make a profit.

We briefly introduce the models of Kyle (1985) [26]. In his equilibrium model, he make following assumptions.

1. Three kinds of traders in market: a single risk neutral insider, random noise traders, competitive risk neutral market makers,
2. The informed trader is assumed to maximize expected profits,
3. the noise traders submit their order randomly,
4. Market makers determine market prices equal to the expectation of the liquidation value of the commodity,

Under these conditions first, Kyle (1985) [26], investigate a single auction model. Kyle (1985) [26]'s single auction model involves 5 kinds of random variables,

1.  $\tilde{\nu}$ : the *ex post* liquidation value of the risky asset,  $\tilde{\nu} \sim N(p_0, \Sigma_0, )$ ,
2.  $\tilde{\mu}$ : the quantity traded by by noise traders,  $\tilde{\mu} \sim N(0, \sigma_\mu^2)$ ,
3.  $\tilde{\chi}$ : the quantity traded by the insider,
4.  $\tilde{p}$ : the market price,
5.  $\tilde{\pi}$ : the profit of the insider,  $\tilde{\pi} = (\tilde{\nu} - \tilde{p})\tilde{\chi}$ ,

## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

and, 2 kinds of measurable functions

1.  $X$  : the trading strategy of informed traders,  $\tilde{\chi} = X(\tilde{\nu})$ ,
2.  $P$  : pricing rule of market makers,  $\tilde{p} = P(\tilde{\chi} + \tilde{\mu})$ .

Based on these settings, an equilibrium is defined as a pair  $X, P$  satisfy following optimization problem.

$$\begin{aligned} & \underset{X, P}{\text{maximize}} && E[\tilde{\pi}(X, P) | \tilde{\nu} = \nu] \\ & \text{subject to} && \tilde{p}(X, P) = E[\tilde{\nu} | \tilde{\chi} + \tilde{\nu}]. \end{aligned} \tag{2.1.1}$$

Kyle (1985) [26] shows that there exists the unique equilibrium (that is, there exists the unique solution of the optimization problem (2.1.1)) under the linearity assumptions on the trading strategy of informed traders,  $X$ , pricing rule of market makers,  $P$ .

**Theorem 2.1.1.** *(Kyle 1985) There exists the unique equilibrium in which  $X$  and  $P$  are linear functions. Defining constants  $\beta$  and  $\lambda$  by  $\beta := (\sigma_\mu^2 / \Sigma_0)^{1/2}$  and  $\lambda := 2(\sigma_\mu^2 / \Sigma_0)^{-1/2}$ , the equilibrium  $P$  and  $X$  are given by*

$$X(\tilde{\nu}) = \beta(\nu - p_0), \quad P(\tilde{\chi} + \tilde{\nu}) = p_0 + \lambda(\tilde{\chi} + \tilde{\nu}).$$

*Proof.* Suppose that for constants  $\alpha, \beta, \gamma, \lambda$ , linear functions  $P$  and  $X$  are given by

$$P(y) = \gamma + \lambda y, \quad X(\nu) = \alpha + \beta \nu$$

Since  $P$  is linear, profits can be written

$$E\{[\tilde{\nu} - P(x + \tilde{\mu})]x | \tilde{\nu} = \nu\} = (\nu - \mu - \lambda x)x.$$

Maximizing the quadratic objective function(w.r.t  $x$ ) yields

$$1/\beta = 2\lambda, \quad \alpha = -\mu\beta \tag{2.1.2}$$

Since  $X$  and  $P$  are linear, the market efficiency condition is

$$\mu + \lambda y = E\{\tilde{\nu} | \alpha + \beta \tilde{\nu} + \tilde{\mu} = y\}. \tag{2.1.3}$$

## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

By the normality, the regression is linear so that we can get

$$\lambda = \frac{\beta \Sigma_0}{\beta^2 \Sigma_0 + \sigma_\mu^2}, \quad \mu = p_0 - \lambda(\alpha + \beta p_0). \quad (2.1.4)$$

By (2) , (4) subject to the second order condition  $\lambda > 0$  we can get the results.  $\square$

In this model, the expected profits of the insider are given by  $E(\tilde{\pi}) = \frac{1}{2}(\Sigma_0^2 \sigma_\mu^2)^{1/2}$ . Kyle (1985) [26] then, expands his single auction model to sequential and continuous auction model. Kyle (1985) [26] shows that under the some linearity conditions, the sequential auction and continuous auction model have the unique equilibrium and during the procedures of reaching the unique equilibrium, the informed trader can make his or her profits.

Since Kyle (1985) [26], several theoretical and experimental studies expand on Kyle (1985) [26] to overcome the restrictiveness of the model's single informed investor assumption (Admati and Pfleiderer (1988) [1], Holden and Subrahmanyam (1992, 1994) [22],[23], and Foster and Viswanathan (1993, 1996) [15],[16]). In particular, whereas Kyle (1985) [26]' model shows that an informed investor trades in a gradual manner when there is only one informed investor in the market, Holden and Subrahmanyam (1992) [22] show that informed investors trade aggressively when the market has more than one informed investor, that is, when there is competition among informed investors. Holden and Subrahmanyam (1992) [22] show that, as the number of informed traders in the market increases, so does the informational efficiency of the price, as measured by the conditional variance of the equilibrium market price in each trading period. (See Figure 2.1, which is reproduced from Holden and Subrahmanyam (1992) [22]). That is, the more informed are the investors in the market, the more quickly the price reflects the information.

Schnitzlein (2002) [33]'s and Bossaerts, Frydman, and Ledyard (2014) [5]'s experimental studies show a positive correlation between the number of informed investors and the speed of the price adjustment process. In particular,

## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

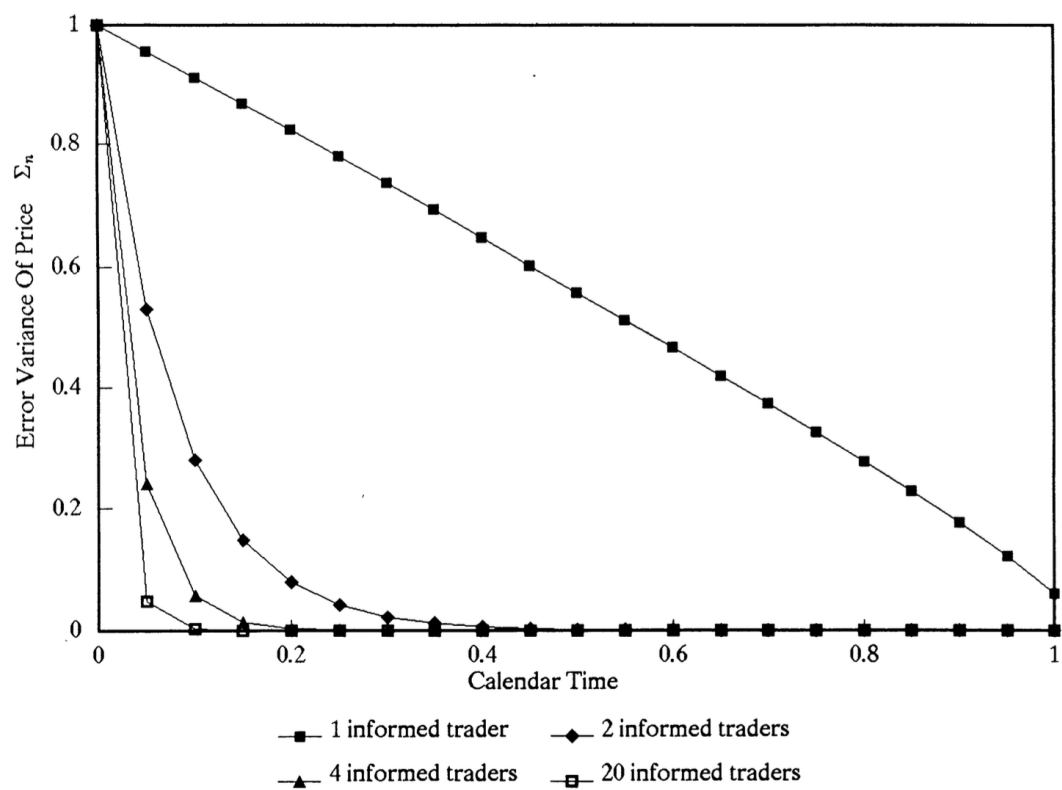


Figure 2.1



## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

Bossaerts et al. (2014) [5], in their experimental study show that the theoretical results of Holden and Subrahmanyam (1992) [22], which are described in Figure 2.1, also hold in their artificial market. Bossaerts et al. (2014) [5]'s empirical results are summarized in Figure 2.2, which is reproduced from Bossaerts et al. (2014) [5].

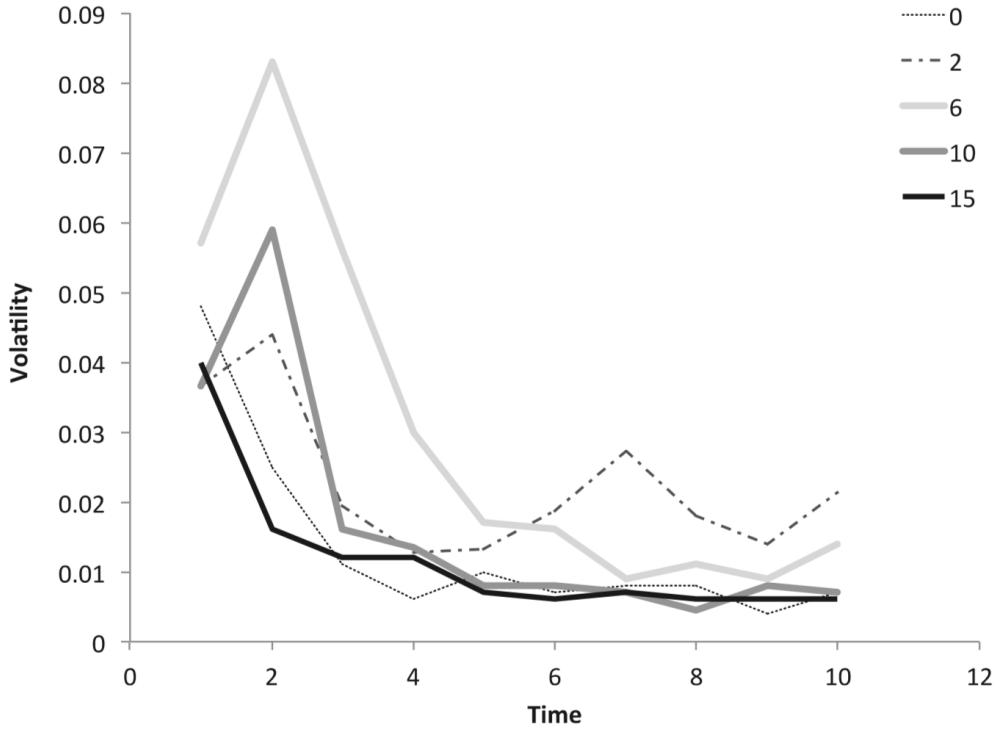


Figure 2.2: Average volatility (measured as transaction price range) per sub-period of 30s, stratified by number of insiders (0, 2, 6, 10, and 15, out of 20 participants).

## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

From these studies on information theory, we can deduce that the more the number of traders in the market who have similar information, the higher the probability that the information is already fully or mostly reflected in the market price and the less valuable the information. Therefore, if we assume that investors in the same position (sell or buy) have similar information, it is likely that, the smaller the investor group, the more valuable its information, supporting motivation.

### **2.1.2 Motivation based on empirical studies of traders' trading skills**

The second motivation of our model is based on the many empirical studies of traders' trading skills. While these studies' results often differ, most of them share the conclusion that there is only a small subset of skilled investors who can beat the market

First, in many empirical studies about skill of fund managers, although there is some controversy about fund managers' market-timing ability, many empirical studies on the skills of fund managers agree that only a small subset of fund managers have sufficient stock-picking ability to ensure superior performance. Second, there is rising interest in the field of investment behavior and the performance of individual traders, which Campell (2006) [9] defines as "household finance," and many of the empirical studies about individual traders' trading skills agree that only a small subset of individual investors are particularly skilled. From these empirical studies, we can conclude that scarcity of skill is a necessary condition for a trading skill to work well in the market, as we often witness in the financial market when a trading strategy works well when it is first introduced and is used by only a small number of traders but is no longer effective when it becomes widely known. A typical example of this phenomenon is high-frequency trading (HFT). According to Clive Cookson (2013) [12], as the HFT strategies became more widely used, making a profit using the strategy became more difficult. According to an es-

## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

timate from Frederi Viens of Purdue University, profit from HFT in the US declined from an estimated peak of \$5bn in 2009 to about \$1.25bn in 2012. Given the importance of the scarcity of trading skill, then if investors in the same position (sell or buy) have similar trading skills (or equivalently, stand on their position based on similar trading skills), it is likely that, the smaller an investor group, the more likely that its trading skills will make the right decision. From this point of view, another motivation of our model is as follows.

**Motivation 2.** *(A motivation based on on trading skills of trader.)*

*The probability that a trading group (a seller or buyer group) will make the right decision based on its trading skills can be measured ex ante by the ratio between the sizes of the two groups.*

## 2.2 The model description

Based on motivations 1 and 2, we first define our key variable, the *sell-buy ratio* ( $SBR$ ). Then we use the variable to develop two hypotheses that describe the relationship between the stock price process and the dynamics of traders' behaviors.

**Definition 2.2.1.** Let  $S_{i,t}(B_{i,t})$  be the number of net sellers (buyers) of a stock  $i$  at time  $t$ . Then, the “Sell-Buy Ratio” (SBR) of the stock  $i$  at time  $t$  is defined by

$$SBR_{i,t} = \frac{S_{i,t}}{B_{i,t}}, \quad (2.2.1)$$

and the natural logarithm of the  $SBR_{i,t}$  is denoted by  $N_{i,t} = \log(SBR_{i,t})$ .

### 2.2.1 SDE form

Utilizing the definition above, we formulate our stock price process model reflecting the dynamics of traders via SDE.

**Hypothesis 1.**

$$dS_t = (\mu + f(N_t))S_t dt + \sigma_1 S_t dW_1 \quad (2.2.2)$$

## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

where,  $f$  is a real-valued function satisfying following three properties.

1. the function  $f$  pass through the origin.
2. the function  $f$  is symmetric about the origin.
3. the function  $f$  is increasing.

The stock price process model given in the *Hypothesis 1*. reflects our two motivations in section 2.1:the *Hypothesis 1* says that,

**Implication 1.** among the seller group and the buyer group,

- (i) the stock price moves to the direction in favor of minority group, not majority group, and
- (ii) the smaller the minor group is, the larger the change of prices is.

**Hypothesis 2.**

$$dN_t = g(N_t)dt + \sigma_2 dW_2 \quad (2.2.3)$$

where,  $g$  is a real-valued function satisfying following three properties.

1. the function  $g$  pass through the origin.
2. the function  $g$  is symmetric about the origin.
3. the function  $g$  is decreasing.

The *Hypothesis 2* says that, in the process of the stock price moving to the direction in favor of minority group, at the same time, the value of  $SBR_t$  changes to the direction of '1', in which the number of net sellers and net buyers are equal. That is, the *Hypothesis 2* implies that

**Implication 2.** among the seller group and the buyer group,

- (i) traders follow (herd to) the behavior of the minority, and
- (ii) the smaller the minor group is, the larger the herding occurs.

## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

The *Hypothesis 2* can be explained by two kinds of point of views, first, by the view of information of a trader group, second, by view of trading skills of trader group. First, information cascading can explain the *Hypothesis 2*. At the early stage of positive (negative) information generation, it is likely that only the small number of investor can know the information. Therefore, SBR is larger (smaller) than 1 because only the small informed investors bet on the buy (sell) position. However, as time goes by, the information cascade over the market so that more investors become informed. Therefore, more informed investors bet on the buy (sell) position which make SBR smaller (larger). Second, skilled traders are highly likely to be not followers but leaders to a price movement in a market because only by a preemptive action (buy/sell), they can make them to beat the market. Therefore, after only a small number of skilled traders bet on buy (sell) position in a market which means SBR is larger (smaller) than 1, many other followers are likely to come to the market with same position with the skilled traders. The new participation of followers to the skilled trader in the market makes SBR smaller (larger) than its previous value.

### 2.2.2 Closed-form solution of the SDE (linear assumption)

Our SDE models have closed-solutions if we assume the linearity in both functions,  $f$  and  $g$ . Under the linearity assumption, we can reformulate the *Hypothesis 1*, and the *Hypothesis 2* as follows.

**Hypothesis 3.** (*Hypothesis 1 with linearity assumption*)

$$dS_t = (\mu + aN_t)S_t dt + \sigma_1 S_t dW_1, \quad (2.2.4)$$

where  $a > 0$ .

**Hypothesis 4.** (*Hypothesis 2 with linearity assumption*)

$$dN_t = -bN_t dt + \sigma_2 dW_2, \quad (2.2.5)$$

## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

where  $b > 0$ .

Under these linearity assumptions, we can deduce the closed-forms of two SDE's, (2.2.4) and (2.2.5).

**Theorem 2.2.2.** *The process*

$$N_t = N_0 e^{-bt} + e^{-bt} \int_0^t e^{bs} \sigma_2 dW(s)_2 \quad (2.2.6)$$

solves the SDE (2.2.5)

*Proof.*

$$d(e^{bt} N_t) = e^{bt} dN_t + b e^{bt} N_t dt = e^{bt} \sigma_2 dW_2 \quad (2.2.7)$$

by integrating both sides of (2.2.7), we can get

$$e^{bs} N_s \Big|_0^t = \int_0^t d(e^{bs} N_s) = \int_0^t e^{bs} \sigma_2 dW(s)_2 \quad (2.2.8)$$

Therefore, we can get the equation

$$e^{bt} N_t - N_0 = \int_0^t e^{bs} \sigma_2 dW(s)_2, \quad (2.2.9)$$

so finally get the desired equation

$$N_t = N_0 e^{-bt} + e^{-bt} \int_0^t e^{bs} \sigma_2 dW(s)_2. \quad (2.2.10)$$

□

**Theorem 2.2.3.** *The process*

$$S_t = S_0 \exp \left[ \int_0^t \left\{ \mu - \frac{1}{2} \sigma_1^2 + a(N_0 e^{-b\tau} + e^{-b\tau} \int_0^\tau e^{bs} \sigma_2 dW_s) \right\} d\tau + \int_0^t \sigma_1 dW_1 \right] \quad (2.2.11)$$

solves the SDE (2.2.4)

## CHAPTER 2. THE STOCK PRICE PROCESS MODEL REFLECTING THE DYNAMICS OF TRADERS' BEHAVIOR

*Proof.* If we divide both sides of (2.2.4) by  $S$ , we can get

$$\frac{dS}{S} = (\mu + aN_t)dt + \sigma_1 dW_1. \quad (2.2.12)$$

Then,

$$d \ln(S_t) = \frac{dS}{S} - \frac{1}{2} \frac{1}{S^2} \sigma_1^2 S_t^2 dt \quad (2.2.13)$$

$$= (r + aN_t - \frac{1}{2} \sigma_1^2)dt + \sigma_1 dW_1, \quad (2.2.14)$$

where the first equality holds by *Ito's Lemma*, and the second equality holds by (2.2.12). By integrating the above equations, we can get

$$\ln(S_s) \Big|_0^t = \ln(S_t/S_0) \quad (2.2.15)$$

$$= \int_0^t (r + aN_s - \frac{1}{2} \sigma_1^2) ds + \int_0^t \sigma_1 dW(s)_1. \quad (2.2.16)$$

$\therefore$

$$S_t = S_0 \exp \left\{ \int_0^t (r + aN_s - \frac{1}{2} \sigma_1^2) ds + \int_0^t \sigma_1 dW(s)_1 \right\}. \quad (2.2.17)$$

By substituting the  $N_s$  in (2.2.17) with (2.2.6), we can get

$$S_t = S_0 \exp \left[ \int_0^t \left\{ \mu - \frac{1}{2} \sigma_1^2 + a(N_0 e^{-b\tau} + e^{-b\tau} \int_0^\tau e^{bs} \sigma_2 dW_s) \right\} d\tau + \int_0^t \sigma_1 dW_1 \right]. \quad (2.2.18)$$

□

# Chapter 3

## Empirical Results

In this chapter, to verify the validity of our model, we perform several empirical tests. To this end, we first describe our proprietary data set. we then, report our empirical results.

### 3.1 Data description

Our proprietary data is composed of the intraday transaction data for all stock (2,131 stocks) listed on the KRX from February 1, 2008 through 30, December 2009 (479 trading days). It includes 822,933 observations (day and stock) from 2,131 stocks . The transaction data include many information about every order and trade occurred in KRX during the periods. The information not only contains the price and quantity of stocks sold or bought, and trade time in milliseconds, but also contains symbols that makes it possible to identify an account for each trade uniquely. With aid of these symbols, we could calibrate the number of sellers and buyers for each observation so that calculate our main measure, sell-buy ratio(SBR).

Panel A of table 1 shows the descriptive statistics of SBR for our sample. <sup>1</sup> The mean of SBR is smaller than 1, 0.94, which is statistically different

---

<sup>1</sup>To remove observation having extremely large SBR, we winsorize sample at 0.1 % and



### CHAPTER 3. EMPIRICAL RESULTS

from 1.<sup>2</sup> It can mean that there are relatively more number of buyers than the number of sellers on average during the sample period. Although mean of SBR can present useful information about our sample, mean of SBR can produce biased result: since SBR is defined on positive domain, and has a shape of right skewness, the distribution of SBR is not symmetric relative to 1. But, as stated above, 1 is critical value so that whether SBR is larger or smaller than 1 is of great interest. Therefore, in order the mean to have a useful information about our sample, it is more desirable to transform the sample to have a symmetric form relative to some number. The one of simplest way of performing this task is just to take natural logarithm for SBR which can make SBR symmetric relative to 0. The mean of SBR taken by natural logarithm is -0.19 and t-value is -344.17. Since 0 of SBR taken by natural logarithm corresponds to 1 of SBR, it implies that information and trading skill of seller group is slightly scarcer than that of buyer group on average during the sample periods, which is consistent result with implication of average SBR.

Panel B of Table 1 shows the sample composition ratio. 32.78 % of total sample have SBR larger than 1, 65.16% of total sample have SBR smaller than 1, and 2.06 % of total sample have SBR equal to 1. Therefore, there are much more sample with SBR smaller than 1 (sample in which seller group have scarcer information and trading skills than buyer group) than sample with SBR larger than 1 (sample in which buyer group have scarcer information and trading skills than seller group). It consistent with the implications of average SBR and average SBR taken by SBR.

Table 2 shows the descriptive statistics of daily returns for all sample stock of our study. The mean of daily returns is positive 0.03 % with t-statistics 7.06, which implies that the positiveness is statistically meaningful.

---

99.9%.

<sup>2</sup>t-statistics of one sample t test for new sample constructed by subtracting '1' from original sample is -105.74. It means that original sample is statistically different from 1 in very strong degree.

CHAPTER 3. EMPIRICAL RESULTS

Panel A											
	min	P1	P5	P10	P25	Median	P75	P90	P95	P99	Max
SBR	0.09	0.22	0.36	0.44	0.61	0.83	1.12	1.51	1.87	2.98	6.00
	Mean	Std	Log mean								
SBR	0.94	0.53	-0.19 (-344.17)								
Panel B											
SBR	Number of Observation						Number of Observation (%)				
SBR > 1	269751						32.78				
SBR = 1	16921						2.06				
SBR < 1	536261						65.16				
Total	822933						100				

Table 1

	min	P1	P5	P10	P25	Median	P75	P90	P95	Max
<i>Daily return (%)</i>	-15.00	-14.23	-6.75	-4.39	-1.85	0.02	1.82	4.54	7.00	15.00
	Mean	Std								
<i>Daily return (%)</i>	0.03	4.32								
	(7.02)									

Table 2

## 3.2 Empirical results

To test Hypothesis 1, we first investigate the relationship between the SBR of stocks on a day and their next-day return. If Hypothesis 1 holds in the financial market, there must be strong positive relationship between the SBR and the next-day return. We then test Hypothesis 2 to determine whether it holds in the financial market.

### 3.2.1 The dynamics of stock price processes

#### 3.2.1.1 Cardinal property of SBR

In this section, we test empirically the first half of Hypothesis 1, the (i) of **Implication** 1, which proposes that the stock price moves in favor of the minority group, not the majority group. To investigate the (i) of **Implication** 1, we investigate the relationship between the cardinal property of SBR and future stock returns.

To this end, we first partition our sample into three groups on the basis of 1 as the SBR of the previous day: a group whose previous day's SBR is larger than 1, a group whose previous day's SBR is equal to 1, and a group whose previous day's SBR is less than 1. We then calibrate and average the daily return for all samples in each group, as shown in Table 3. The average daily return of the group whose previous day's SBR is larger (smaller) than 1 is 0.21% (-0.06%), with a t-statistic of 26.11 (-9.73). Therefore, when there are more sellers (buyers) than buyers (sellers) in a market, so the buyers' (sellers') information and trading skills are scarcer than those of the seller (buyer) group, the buyer (seller) group is likely to be the winner in their trades. The average daily return of the group whose previous-day SBR equals 1 is 0.07%, with a t-statistic of 2.43. This empirical result for this group shows that, although the t-value of the group's daily returns is sufficiently large and the positive mean (0.07%) is statistically meaningful, it is small compared to those of the other two groups. Moreover, the t-value of the average daily returns of the group whose previous-day SBR is 1 is much smaller than that the full, as shown in

## CHAPTER 3. EMPIRICAL RESULTS

Table 1. Therefore, the empirical results shown in Table 3 show that, whereas the buyer (seller) group is likely to be the winner when there are more sellers (buyers) than buyers (sellers) in a market, when there are similar numbers of sellers and buyers in a market, the probability that the buyer group will be the winner is slightly larger than the probability that the seller group will be the winner. This empirical result supports the first part of Hypothesis 1, which states that the stock price moves to favor the minority group, not the majority group.

SBR	average daily return (%)
SBR > 1	0.21 (26.11)
SBR = 1	0.07 (2.43)
SBR < 1	-0.06 (-9.73)

Table 3

### 3.2.1.2 Ordinal property of SBR

In this section, we test empirically the remaining half of Hypothesis 1, (ii) of **Implication 1** : the smaller the minority group, the larger the change in prices. To investigate the (ii) of **Implication 1**, we investigate the relation between ordinal property of SBR and future stock returns.

Table 4 shows the average daily returns and cumulative daily returns of decile (value-weighted and equal-weighted) portfolios that are formed by sorting our sample based on the SBR of the previous day. Panel A of table 4 shows the average daily returns of value-weighted decile portfolios. Decile 1 (minimum SBR) is the portfolio of stocks with the lowest SBR on the previous day, and decile 10 (maximum SBR) is the portfolio of stocks with the highest SBR on the previous day. The difference in the value-weighted average

### CHAPTER 3. EMPIRICAL RESULTS

returns between decile 1 and decile 10 is 0.49%, with a t-statistic of 7.46, a difference that is economically and statistically significant at all conventional levels. The relationship between the value-weighted average daily returns and deciles suggests that a positive relationship between value-weighted average daily returns and deciles, as the average daily returns of deciles 1-7 increase from -0.10% to 0.08%. However, going from decile 7 to decile 10, the average daily returns increase more dramatically, from 0.08% to 0.40%. Therefore, the average daily returns have an increasing trend that is statistically significant, as the standard errors show.

Panel A of Table 4 also shows that the standard errors of the decile portfolios do not increase across deciles, as the standard error of decile 10 is the smallest of all deciles, at 0.09%. Therefore, although the average daily return of the decile 10 portfolio is the largest of the 10 deciles, the risk, measured as variation in returns, of the decile 10 portfolio is the smallest. This result suggests that the SBR is not a risk factor of stocks to be priced in a market but a useful indicator that contains information about the stock's future price movements.

Panel A of Table 4 also shows the cumulative return of the value-weighted decile portfolios and an increasing pattern in the cumulative daily returns that is similar to that of the average daily returns, increasing from -44.43% to 505.33%. Moreover, the difference in the cumulative returns between the minimum SBR portfolio and the maximum SBR portfolio is 549.76%, suggesting that, if we had formed a portfolio by shorting the decile 1 portfolio and longing the decile 10 portfolio every day, we could have made a 549.76% return during the sample periods, which is large if we consider that the length of the sample period is only about two years. Therefore, we re confirm that the SBR is a useful indicator that contains information about stocks' future price movements.

Panel B of Table 4 shows a similar but weaker positive correlation between average daily returns (cumulative returns) and the decile portfolios for equal-weighted portfolios. The first column of Panel B of Table 4 shows the average

### CHAPTER 3. EMPIRICAL RESULTS

daily returns of the equal-weighted decile portfolios during the sample periods: While the average daily returns of deciles 1-7 are similar, in the range of 0.05 to 0.16%, the decile 7-10 portfolios' average daily returns increase from 0.16% to 0.49%. The difference in the average daily return between the minimum SBR portfolio and the maximum SBR portfolio is 0.44%, with a t-statistic of 11.60. The standard errors of the decile portfolios also have a minimum values of 0.08% in the maximum SBR portfolio, which is the portfolio with the largest average daily return.

Third column of Panel B of Table 4 presents the cumulative returns of the decile portfolios. The cumulative returns of decile portfolios 1-5 have a zigzag shape in the range of 16.38% to 66.25%. However, the cumulative returns of the decile portfolios 5-10 increase from 66.25% to 865.78%. Moreover, the difference in the cumulative returns between the minimum SBR portfolio and the maximum SBR portfolio is 849.39%, so we could have made a 849.39% return during the sample periods if we had formed a portfolio by shorting portfolio 1 and longing portfolio 10 every day.

These portfolio analyses suggest two important results: First, the daily return has a positive correlation with the SBR of the previous day. Second, the although there is a positive correlation with the next day's daily return, the SBR is not a risk factor. These two empirical results strongly the remaining half of Hypothesis 1, (ii) of **Implication 1**: that the smaller the minority group is, the larger the change in prices.

# CHAPTER 3. EMPIRICAL RESULTS

Decile	Panel. A Value-weighted portfolio				Cumulative Return	
	SBR	Average daily return	Return difference with VWMR	Return	Return difference with VWMR	
1	0.37	-0.10 [0.10]	-0.22	-44.43	-106.61	
2	0.53	-0.04 [0.11]	-0.16	-28.88	-91.06	
3	0.63	0.03 [0.11]	-0.10	-1.68	-63.85	
4	0.72	0.03 [0.12]	-0.09	-0.54	-62.72	
5	0.81	0.05 [0.12]	-0.08	7.19	-54.99	
6	0.89	0.06 [0.11]	-0.06	18.01	-44.17	
7	0.99	0.08 [0.10]	-0.05	26.54	-35.64	
8	1.12	0.14 [0.10]	0.02	71.83	9.66	
9	1.33	0.24 [0.10]	0.12	180.64	118.46	
10	2.04	0.40 [0.09]	0.27	505.33	443.15	
Max- Min t(Max-Min)		0.49 (7.46)		549.76		
VW market return (VW/MR)		0.12		62.18		

Table 4A

# CHAPTER 3. EMPIRICAL RESULTS

Decile	SBR	Panel. B Equal-weighted portfolio			Cumulative Return	
		Average daily return		Return difference with VWMR	Return	Return difference with VWMR
		Return				
1	0.37	0.05 [0.09]	-0.14		16.38	-110.61
2	0.53	0.13 [0.10]	-0.06		64.90	-62.09
3	0.63	0.13 [0.10]	-0.06		65.03	-61.96
4	0.72	0.13 [0.10]	-0.06		63.28	-63.72
5	0.81	0.13 [0.10]	-0.06		66.25	-60.74
6	0.89	0.15 [0.10]	-0.04		82.10	-44.89
7	0.99	0.16 [0.10]	-0.04		89.16	-37.84
8	1.12	0.23 [0.09]	0.03		165.54	38.54
9	1.33	0.33 [0.09]	0.14		347.50	220.51
10	2.04	0.49 [0.08]	0.30		865.78	738.78
Max- Min t(Max-Min)		0.44 (11.60)			849.39	
VW market return (VWMR)		0.19			126.99	

Table 4B



## CHAPTER 3. EMPIRICAL RESULTS

### 3.2.2 The dynamics of SBR

#### 3.2.2.1 Cardinal analysis of the dynamics of traders

In this section, we test the first half of Hypothesis 2, (i) of **Implication 2** empirically: traders follow (herd to) the behavior of the minority. To investigate the (i) of **Implication 2**, we first define the change rate of traders.

**Definition 3.2.1.** (The change rate of traders) Let  $SBR_{t+1}^i$  be a sell-buy ratio of a stock  $i$  in day  $t + 1$ , and  $BSR_{t+1}^i$  be a buy-sell ratio of a stock  $i$  in day  $t + 1$ ,

*the change rate of traders of a stock  $i$  at day  $t =$*

$$\begin{cases} \frac{SBR_{t+1}^i}{SBR_t^i} - 1, & SBR_t^i > 1, \\ \frac{BSR_{t+1}^i}{BSR_t^i} - 1, & SBR_t^i < 1, \\ \log \frac{SBR_{t+1}^i}{SBR_t^i} - 1, & SBR_t^i = 1, \end{cases}$$

Therefore, the *change rate of traders*<sup>3</sup> measures different quantity for three groups.

1. A group with SBR larger than 1 (the number of sellers > the number of buyers) : measures the change rate of ratio of buyers to sellers.
2. A group with SBR smaller than 1 (the number of sellers < the number of buyers): measures the change rate of ratio of sellers to buyers.
3. A group with SBR equal to 1 (the number of sellers = the number of buyers): measures the change rate of ratio of buyers to sellers.

Table 5 presents the average change rate of the traders for each of the three groups. The average change rate of traders for the group whose SBR is larger

---

<sup>3</sup>The reason we do not simply use growth rate of  $SBR (= \frac{SBR_{t+1}^i}{SBR_t^i} - 1)$  for all the three groups is that growth rate of SBR shows asymmetric behavior relative to 1. Therefore, it is necessary to define different formulas for both a group with SBR larger than 1 and a group with SBR smaller than 1 respectively to balance a scale of the measure.

## CHAPTER 3. EMPIRICAL RESULTS

than 1 and the group whose SBR is less than 1 is negative (-0.22 and -0.06, respectively), and the t-statistics are sufficiently large (-268.30, -81.96, respectively). Therefore, when there are more sellers (buyers) than buyers (sellers) in a market, more buyers (sellers) than sellers (buyers) than there were today are likely to come to the market tomorrow. The SBR is likely to decrease (increase) tomorrow if today's SBR is larger (smaller) than 1. The average change rate of traders in the group whose SBR is 1 is -0.16, and the t-statistic is -37.38. Although the t-value of the SBR's growth rate for this group is sufficiently large (-37.38), the negative mean (-0.16) is statistically meaningful, as it is small compared to those of the other groups (-268.30 for the group whose SBR is larger than 1 and -81.96 for the group whose SBR is less than 1). These empirical results support the first half of Hypothesis 2, (i) of **Implication**.

SBR	Change rate of traders
SBR > 1	-0.22 (-268.30)
SBR = 1	-0.16 (-37.38)
SBR < 1	-0.06 (-81.96)

Table 5

### 3.2.2.2 Ordinal analysis of the dynamics of traders

In this section, we test the remaining half of Hypothesis 2 ((ii) of **Implication** 2) empirically: the smaller the minor group is, the larger the herding is. To investigate the (ii) of **Implication** 2, we first develop a herding measure appropriate for our analysis. The most popular herding measure is the herding measure defined by Lakonishok et al. (1992) [27] (LSV herding measure), which quantifies how many fund managers buy or sell a stock compared to its benchmark in a given quarter. For a stock  $i$ , the LSV herding measure is defined in Definition 3.2.2

## CHAPTER 3. EMPIRICAL RESULTS

**Definition 3.2.2.** (LSV Herding measure:based on the number of investors)

$$H(i) = \left| \frac{B(i)}{B(i) + S(i)} - p(i) \right| - AF(i), \quad (3.2.1)$$

where,

$B(i)$ : the number of fund managers (net buyers) who increase their holdings of the stock in the quarter.

$S(i)$ : the number of fund managers (net sellers) who decrease their holdings in the quarter.

$p(t)$ : the expected value of  $\frac{B(i)}{B(i) + S(i)}$  in that quarter.

$AF(i)$ :an adjustment factor that accounts for the null hypothesis in which there is no herding behavior.

Therefore,(3.2.1) can be interpreted as an unsigned difference between the ratio of a stock's buyer,  $\frac{B(i)}{B(i) + S(i)}$ , and its cross-sectional benchmark,  $p(i)$ . Measuring the herding behavior using Lakonishok et al.'s (1992) [27] method has some limitations in applying to our empirical study. Since it aggregates all buyers and sellers without taking each account into consideration, we cannot determine which traders move first and which move later. By exploiting our unique account data, we can identify the accounts for all trades at day  $t$ , so we can construct a herding measure that considers which investors lead and which follow in their trading. That is, we can construct the measure to distinguish new buyers from all buyers. To construct this measure, we adopt the notion of the Markov transition matrix. In so doing, we first partition the investors that traded stock  $i$  from day  $t$  stock market to day  $t + 1$ —denoting these investors as  $Investors_t^{t+1}$ —into three categories at both day  $t$  stock market and day  $t + 1$ , respectively, with 6 groups in total.

At day  $t$ , we partition  $Investors_t^{t+1}$  into categories  $S_t$ ,  $B_t$ , and  $N_t$ , where

$S_t$ : Investors who sell stock  $i$  in the stock market at day  $t$ .

$B_t$ : Investors who buy stock  $i$  in the stock market at day  $t$ .

$N_t$ : Investors who do nothing in the stock market at day  $t$  but sell or buy stock

### CHAPTER 3. EMPIRICAL RESULTS

$i$  at day  $t + 1$ .

At day  $t+1$ , we also partition  $Investors_t^{t+1}$  into three categories,  $S_{t+1}$ ,  $B_{t+1}$ , and  $N_{t+1}$ , where

$S_{t+1}$ : Investors who sell stock  $i$  at day  $t + 1$ .

$B_{t+1}$ : Investors who buy stock  $i$  at day  $t + 1$ .

$N_{t+1}$ : Investors who do nothing at day  $t + 1$ . But sell or buy stock  $i$  in the stock market at day  $t$ .

Based on these definitions, the following relationships hold.

$$Investors_t^{t+1} = S_t \cup B_t \cup N_t = S_{t+1} \cup B_{t+1} \cup N_{t+1}, \quad (3.2.2)$$

$$S_t \cap B_t = \emptyset, B_t \cap N_t = \emptyset, S_t \cap N_t = \emptyset, \quad (3.2.3)$$

$$S_{t+1} \cap B_{t+1} = \emptyset, B_{t+1} \cap N_{t+1} = \emptyset, S_{t+1} \cap N_{t+1} = \emptyset. \quad (3.2.4)$$

Eq. (3.2.2), (3.2.2), and (3.2.2) state that  $Investors_t^{t+1}$  is partitioned into  $S_t, B_t$ , and  $N_t$  and  $S_{t+1}, B_{t+1}$ , and  $N_{t+1}$ , respectively.

In the second step, two stochastic vectors  $V, W$ , and a Markov transition matrix  $M$  are defined as follows:

$$V = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}, W = \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix}, M = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix}, \quad (3.2.5)$$

where  $v_i (i = 1, 2, 3), w_i (i = 1, 2, 3), a_{i,j} (i = 1, 2, 3, \text{ and } j = 1, 2, 3)$  is defined as follows.

$v_1$ : (the number of  $S_t$ ) / (the number of  $Investors_t^{t+1}$ ),

$v_2$ : (the number of  $B_t$ ) / (the number of  $Investors_t^{t+1}$ ),

$v_3$ : (the number of  $N_t$ ) / (the number of  $Investors_t^{t+1}$ ),

$w_1$ : (the number of  $S_{t+1}$ ) / (the number of  $Investors_t^{t+1}$ ),

$w_2$ : (the number of  $B_{t+1}$ ) / (the number of  $Investors_t^{t+1}$ ),

$w_3$ : (the number of  $N_{t+1}$ ) / (the number of  $Investors_t^{t+1}$ ),

$a_{1,1}$ : (the number of  $S_{t+1} \cap S_t$ ) / (the number of  $S_t$ ),

$a_{1,2}$ : (the number of  $S_{t+1} \cap B_t$ ) / (the number of  $B_t$ ),

## CHAPTER 3. EMPIRICAL RESULTS

$a_{1,3}$ : (the number of  $S_{t+1} \cap N_t$ ) / (the number of  $N_t$ ),  
 $a_{2,1}$ : (the number of  $B_{t+1} \cap S_t$ ) / (the number of  $S_t$ ),  
 $a_{2,2}$ : (the number of  $B_{t+1} \cap B_t$ ) / (the number of  $B_t$ ),  
 $a_{2,3}$ : (the number of  $B_{t+1} \cap N_t$ ) / (the number of  $N_t$ ),  
 $a_{3,1}$ : (the number of  $N_{t+1} \cap S_t$ ) / (the number of  $S_t$ ),  
 $a_{3,2}$ : (the number of  $N_{t+1} \cap B_t$ ) / (the number of  $B_t$ ),  
 $a_{3,3}$ : (the number of  $N_{t+1} \cap N_t$ ) / (the number of  $N_t$ ).

In this setting, we can deduce an important relation between  $V, W, M$  by direct calculation.

**Proposition 3.2.3.** *For stochastic vectors  $V$  and  $W$ , and matrix  $M$  defined in 3.2.5, the following equation holds.*

$$W = MV. \quad (3.2.6)$$

*Proof.* The result follows from the direct calculation.  $\square$

Using the two vectors  $V$  and  $W$  and a Markov transition matrix  $M$ , we can capture changes in investor trading behavior for stock  $i$  between day  $t$  stock market and day  $t + 1$ . The value of  $a_{2,3} \times v_3$  represents the ratio of new buyers relative to  $Investors_t^{t+1}$  at day  $t + 1$ . However, although  $a_{2,3} \times v_3$  quantifies the ratio of new buyers, it is affected by the ratio of buyers in day  $t + 1$ ,  $w_2$ , because  $a_{2,3} \times v_3$  is one of the three components of  $w_2 (= a_{2,1} \times v_1 + a_{2,2} \times v_2 + a_{2,3} \times v_3)$ . Therefore, if the ratio of buyers in day  $t + 1, w_2$ , is large (small), then  $a_{2,3} \times v_3$  is likely to be large (small). Because of this effect, we must control the ratio of buyers,  $w_2$ , in  $a_{2,3} \times v_3$  by constructing a Markov Transition Matrix herding measure. On this setting, We define a new herding measure as follow.

**Definition 3.2.4.** (The Markov Transition Matrix herding measure)

For a stock  $i$ , the Markov Transition Matrix herding measure  $H_{MTM}(i)$  at day  $t$  is defined as follows.

$$H_{MTM}(i) = \frac{a_{2,3} \times v_3}{w_2}. \quad (3.2.7)$$

Table 6, which shows the results from the analysis of the average Markov Transition Matrix herding measure,  $H_{MTM}$ , for ten groups. Table 6 shows

## CHAPTER 3. EMPIRICAL RESULTS

that the averages of  $H_{MTM}$  also show generally increasing trends along the ten groups (except decile 1), in range of 69.80 percent to 72.69 percent, and the increasing trend is statistically meaningful, as the standard errors of the averages show. This empirical results strongly supports the remaining half of Hypothesis 2, (ii) of **Implication 2**: the smaller the minor group is, the larger the herding occurs.

<i>Average <math>H_{MTM}</math> at day <math>t+1</math></i>	
Group	$H_{MTM}(\%)$
1	71.1049
(Min <i>IB</i> )	(0.1173)
2	69.8039
	(0.1154)
3	69.9837
	(0.1102)
4	70.0391
	(0.1075)
5	70.3204
	(0.1047)
6	70.2264
	(0.1051)
7	70.8610
	(0.1127)
8	71.7033
	(0.1086)
9	72.2677
	(0.1154)
10	72.6895
(Max <i>IB</i> )	(0.1243)

Table 6

### 3.3 Robustness check: subperiod test

To test whether our results also hold for different sub-periods, we construct eight sub-periods of approximately the same length. The first seven sub-periods last three months, and the last sub-period lasts two months. Table 7 shows that our main results are robust for all eight sub-periods: average stock return of day  $t + 1$  is positively correlated with SBR in all eight sub-periods, suggesting

## CHAPTER 3. EMPIRICAL RESULTS

that our stock price process model suggested in *Hypothesis 1* explain real financial market well.

# CHAPTER 3. EMPIRICAL RESULTS

<i>Sub period test:</i>											
<i>Average stock return at day t+1</i>											
Group	20080201~ 20080430	20080502~ 20080731	20080801~ 20081031	20081103~ 20090130	20090202~ 20090430	20090504~ 20090731	20090803~ 20091030	20091102~ 20091229			
1	0.0402	-0.2401	-1.0804	0.2405	0.7011	0.2663	0.0186	0.1937			
(Min SBR)	(0.0375)	(0.0375)	(0.0628)	(0.0568)	(0.0477)	(0.0366)	(0.0340)	(0.0430)			
2	0.2012	-0.1885	-0.8845	0.4197	0.7909	0.1913	0.0839	0.2386			
(0.0370)	(0.0381)	(0.0381)	(0.0629)	(0.0571)	(0.0473)	(0.0378)	(0.0333)	(0.0436)			
3	0.2015	-0.1739	-0.7085	0.3637	0.7370	0.2522	0.0687	0.2905			
(0.0387)	(0.0384)	(0.0384)	(0.0642)	(0.0586)	(0.0487)	(0.0398)	(0.0348)	(0.0446)			
4	0.2106	-0.1747	-0.7371	0.3653	0.6486	0.2043	-0.0126	0.3007			
(0.0384)	(0.0406)	(0.0406)	(0.0647)	(0.0581)	(0.0503)	(0.0407)	(0.0375)	(0.0475)			
5	0.2012	-0.2146	-0.6823	0.3716	0.6798	0.2321	-0.0286	0.2818			
(0.0396)	(0.0412)	(0.0412)	(0.0662)	(0.0597)	(0.0506)	(0.0427)	(0.0390)	(0.0480)			
6	0.2276	-0.0321	-0.5708	0.3448	0.6845	0.1687	0.0166	0.2354			
(0.0408)	(0.0443)	(0.0443)	(0.0653)	(0.0594)	(0.0510)	(0.0443)	(0.0412)	(0.0493)			
7	0.2012	-0.0815	-0.4297	0.4846	0.7429	0.1334	-0.0046	0.2390			
(0.0415)	(0.0445)	(0.0445)	(0.0648)	(0.0587)	(0.0507)	(0.0452)	(0.0416)	(0.0510)			
8	0.2858	0.0086	-0.4294	0.5007	0.7950	0.2240	0.0198	0.3263			
(0.0409)	(0.0439)	(0.0439)	(0.0633)	(0.0585)	(0.0498)	(0.0441)	(0.0412)	(0.0499)			
9	0.4214	0.0625	-0.2570	0.6057	0.9477	0.4444	0.1971	0.3157			
(0.0398)	(0.0424)	(0.0424)	(0.0610)	(0.0567)	(0.0489)	(0.0415)	(0.0390)	(0.0458)			
10	0.5729	0.2453	0.0568	0.8350	1.0930	0.6529	0.4075	0.4788			
(Max SBR)	(0.0375)	(0.0381)	(0.0566)	(0.0560)	(0.0471)	(0.0387)	(0.0364)	(0.0412)			

Table 7



## Chapter 4

# Return prediction via a machine learning technique

Recently, there are many trials to predict stock returns via various machine learning techniques. Enke, and Thawornwong (2005) [14] use neural network models for level estimation and classification. They show that the trading strategies given by the classification models generate higher risk-adjusted profits than the buy-and-hold strategy, as well as those given by the level-estimation based on forecasts of the neural network and linear regression models. Huang, Nakamori, Wang (2005) [24] compare the prediction performance of Support vector machine (SVM) with those of Linear Discriminant Analysis, Quadratic Discriminant Analysis and Elman Backpropagation Neural Networks. They show that SVM outperforms the other classification methods. Nguyen, Shirai, and Velcin (2015) [28] suggest a model to predict stock return using the sentiment from social media. Their model shows the better accuracy performance than the models using historical prices only. Patel et al. (2015) [31] compares four prediction models, Artificial Neural Network (ANN), Support Vector Machine (SVM), random forest and naive-Bayes in predicting stock price returns. they show that random forest outperforms other three prediction models on overall performance.

Most of these line of studies utilize machine learning techniques to directly

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

predict stock price return. In this paper, we have shown that the SBR (sell-buy ratio) has the huge prediction power on stock returns. utilizing this fact, in this chapter, we do not predict stock return directly, but by predicting SBR, predict stock return indirectly.

### 4.1 Test data set description

We have acquired a proprietary dataset from KOSKOM to investigate trading patterns of the investors during the most recent (three-year) period. Specifically, our dataset contains details of all transactions occurred from June 1, 2013 through May 31 2016 (736 trading days). The transaction data include the date and time of the transaction, a stock identifier, trader type (which classifies the seller or the buyer into three groups: domestic individuals, domestic institutions, or foreign institutions.)

### 4.2 Data filtration

Our dataset provides precise information about the specific type of trader for either side of any executed trade. For example, we can identify whether the seller (or the buyer) is a domestic individual investor or a domestic institutional investor or a foreign institutional investor. To predict SBR, we mainly utilize these kinds of information. In order to fully exploit the information, we take several filtrations. (the filtrations described in this section is same with our prior study, Chay and Kim (2017) [10]) Our filters are based on two layers: the first set of filters is implemented at the stock-day level, and the second set at the stock level. Filters at the stock-day level are as follows. First, we exclude any stock-day observation (on day  $t$ ) that shows a market capitalization less than 300 billion Korean Won (smaller than 300 million U.S. dollars) at the end of day  $t - 1$ . Second, we exclude any stock-day observation with the stock price less than 5,000 Won (less than US \$ 5) at the end of day  $t - 1$ . Third, we exclude any stock-day observation that records trading frequency less than 40 times during day  $t$ . Fourth, we exclude any stock-day observation whose trading

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

volume is less than 1,000 shares during day  $t$ . Next, we impose a stock-level filter on the sample already screened by the above filters. We remove the stocks that were traded for less than 400 trading days out of the entire 736 trading days of our sample period. Our filtering process produced a sample of 357 distinct stocks with 238,462 stock-day observations. The number of stocks on each trading day varies between 280 to 354. Table 8 reports summary statistics for our final sample.

Summary statistics.

	N	Average of:			
		Size (₩ BM)	Close price /day	Trades /day	Trading volume/day
All firms	357	3543.04	102107.41	7308.11	413630.24
Smallest tertile	119	467.79	45540.09	3469.12	153971.41
Middle tertile	119	1049.57	74382.22	5824.41	251973.21
Largest tertile	119	8375.70	181192.64	11912.77	441281.56

Table 8

## 4.3 Key predictors

### 4.3.1 Interaction between types

Our data provide the records that allow us to identify the exact investor type (i.e., group) of both the buy-side and sell-side traders of a transaction. This feature of the data allows us to measure the volume of trading occurring between and within investor groups. In Table 9, we quantify the average proportion of trading by each investor group in each stock and also the trading interaction among investor groups. Table 9, report the average proportion of trading volume by individuals (households), institutions, and foreigners. For each stock, we first take the time-series average of daily relative trading volume by each investor group. Then, we calculate and report the cross-sectional averages across

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

all stocks in the entire sample or in each size tertile. We observe two facts from Table 9. First, across all stocks, individuals account for the largest proportion of trading: individuals' average trading volume explains 51.01% of all trading, while institutions and foreigners engage in only 21.51% and 19.83%, respectively. The remaining proportions of total trading are related to trades executed by government, non-financial institutions, and foreign individuals. Second, institutions and foreigners tend to focus more on trading large stocks than small stocks. In the smallest tertile, institutions and foreigners are involved in only 16.79% and 13.69% of trading, respectively, as compared to 62.62% represented by individuals. In contrast, individuals' presence diminishes significantly in the largest tertile, accounting for only 36.76% of trading. Instead, institutions and foreigners display much higher proportional trading of 28.84% and 25.51%, respectively, comparable to individuals' trading.

Table 10 reports interaction among three investor groups: individuals, institutions, and foreigners. It also reports within-group trading activities. Based on the three investor groups, we form a 3x3 matrix and report proportional trading in the lower triangle elements. Diagonal elements represent trading within each group. Off-diagonal elements represent trading interaction between two different investor types. We first calculate actual proportion of trading volume executed between and within investor groups on each day for each stock. We then calculate the time series average for each stock. The figures reported in Table 10 are the cross-sectional averages of the time-series averages across stocks in each element of the matrices representing trading pairs within and between investor groups, together with the corresponding standard errors. Focusing on the figures of the diagonal elements that represent trading within each group, we find substantial amounts of trading between two individuals. In the whole sample, 32.31% of trading occurred between individuals. In the smallest tertile, we find nearly half of trading (45.03%) is accounted for by trading between two individuals. By contrast, only 17.11% trading is made between two individuals in the largest tertile, implying that institutional and foreign investors are more active in this tertile.

# CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

	Percentage of trading		
	Households	Institutions	Foreigners
All firms	51.01 (1.03)	21.51 (0.52)	19.83 (0.53)
Smallest tertile	62.62 (1.51)	16.79 (0.90)	13.69 (0.50)
Largest tertile	36.76 (1.35)	25.51 (0.59)	28.84 (0.93)

Table 9

# CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

	Interaction		
	Households	Institutions	Foreigners
<i>All firms</i>			
Households	32.31 (1.14)		
Institutions	20.29 (0.39)	5.85 (0.22)	
Foreigners	17.11 (0.22)	11.03 (0.39)	5.76 (0.31)
<i>Smallest tertile</i>			
Households	45.03 (1.89)		
Institutions	19.25 (0.81)	4.01 (0.35)	
Foreigners	15.93 (0.38)	6.30 (0.45)	2.58 (0.19)
<i>Largest tertile</i>			
Households	17.36 (1.22)		
Institutions	19.86 (0.52)	7.10 (0.30)	
Foreigners	18.95 (0.32)	16.95 (0.61)	10.89 (0.67)

Table 10

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

Based on the filtered sample, we define first kinds of key predictors as follows.

**Definition 4.3.1.** [Interactions between types] Let  $H, I, F$  indicate the total trading volume of stock  $i$  on day  $t$  by households (individuals), institutions, and foreigners, respectively.

- $[A, B]$ : the ratio of volume of trading between buyers of investor type A and sellers of investor type B relative to the total trading volume of stock  $i$  on day  $t$ ;
- $[\tilde{A}, B]$ : the ratio of volume of trading between buyers of investor type A who initiate their trades and sellers of investor type B relative to the total trading volume of stock  $i$  on day  $t$ ;
- $[A, \tilde{B}]$ : the ratio of volume of trading between buyers of investor type A and sellers of investor type B who initiate their trades relative to the total trading volume of stock  $i$  on day  $t$ ;

The definition 4.3.1 induces 27 variables measuring interaction between types:

**Predictors 1.** (*Interaction variables*)

1.  $[H, H], [H, I], [H, F], [I, H], [I, I], [I, F], [F, H], [F, I], [F, F]$ .
2.  $[\tilde{H}, H], [\tilde{H}, I], [\tilde{H}, F], [\tilde{I}, H], [\tilde{I}, I], [\tilde{I}, F], [\tilde{F}, H], [\tilde{F}, I], [\tilde{F}, F]$ .
3.  $[H, \tilde{H}], [H, \tilde{I}], [H, \tilde{F}], [I, \tilde{H}], [I, \tilde{I}], [I, \tilde{F}], [F, \tilde{H}], [F, \tilde{I}], [F, \tilde{F}]$ .

*every predictor variables are calibrated in daily horizon.*

### 4.3.2 LSV herding measure of each types

Although the original LSV herding measure, which is described in Definition 3.2.2, use account data of traders, in many cases, it is difficult or impossible to access the account data of traders. Therefore, many researchers use the

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

modified LSV herding measure which use the number of trades of traders for proxy of the number of traders (Choe, Kho and Stulz (1999) [?]; Zhou and Lai (2009) [?] ; Venezia et al. (2011) [36]; Hsieh (2013) [?]). The modified LSV herding measure is defined as follows.

**Definition 4.3.2.** (LSV Herding measure:based on the number of trades of investors) For stock  $i$  on day  $t$ , the *herding measure* of type A investors is defined as

$$Herd_{i,t}(A) = \left| \frac{B_{i,t}(A)}{N_{i,t}} - \frac{B_t(A)}{N_t(A)} \right| - E_X \left[ \left| \frac{X}{N_{i,t}} - \frac{B_t(A)}{N_t(A)} \right| \right], \quad (4.3.1)$$

where,

$B_{i,t}(A)(S_{i,t}(A))$ : the number of buy (sell) trades executed by type A investors for stock  $i$  on day  $t$ .

$N_{i,t}(A)(= B_{i,t}(A) + S_{i,t}(A))$ : the sum of buy and sell trades executed by type A investors for stock  $i$  on day  $t$ .

$B_t(A)$ : the aggregate buy trades of type A investors across all stocks on day  $t$ .

$N_t(A)$ : the sum of aggregate buy and sell trades of type A investors across all stocks on day  $t$ .

$X$  : a random variable following binomial distribution  $B\left(N_{i,t}, \frac{B_t(A)}{N_t(A)}\right)$ .

The second term in Equation (4.3.1), named as the *adjustment factor* in Lakonishok et al. (1992) [27], can be calculated by the following formula:

$$E_X \left[ \left| \frac{X}{N_{i,t}(A)} - \frac{B_t(A)}{N_t(A)} \right| \right] = \sum_{k=0}^{N_{i,t}(A)} \binom{N_{i,t}(A)}{k} \left( \frac{B_t(A)}{N_t(A)} \right)^k \left( 1 - \frac{B_t(A)}{N_t(A)} \right)^{N_{i,t}(A)-k} \left| \frac{X}{N_{i,t}(A)} - \frac{B_t(A)}{N_t(A)} \right|, \quad (4.3.2)$$

Although there is a simple approximation for Equation (4.3.1) (see Appendix A of Venezia et al. (2011) [36]), we employ the exact formula given



## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

by Equation (4.3.1) in calculating our herding measure. The herding measure,  $Herd_{i,t}(A)$ , is designed to have a value of zero if there is no herding behavior among type A investors in trading stock  $i$  on day  $t$  and have a larger positive value if there is a high degree of herding. For each stock, we calculate daily herding measure for individual investors, institutional investors, and foreign investors separately according to Equation Equation (4.3.1). Descriptive statistics for the time- series averages of daily herding measures are reported in Table 11A. Panel A shows the herding measures for each investor group for the full sample. The mean (median) herding measure is 11.43% (11.32%) for individual investors, 18.79 % (18.15%) for institutional investors, and 17.43% (17.59%) for foreigners. In Panel B and Panel C of Table 11, we show our herding measures after sorting our sample stocks based on the market cap and then assigning them into two groups of 200 stocks each: large and small stocks. As the figures in Panels B and C indicate, individual investors herd more in trading large stocks as compared to their trading in small stocks. In contrast, domestic institutions and foreigners demonstrate much higher degrees of herding when they trade small stocks than when they trade large stocks. Overall our findings suggest that, regardless of the investor type, investors in the same type have strong tendency to trade in the same direction. Institutions and foreigners herd more than individuals. We find that domestic as well as foreign institutions herd more when they trade small stocks than when they trade large stocks. In contrast, individuals tend to herd less when they trade small stocks. LSV herding measure induce 3 predictors.

### **Predictors 2.** (*Herding variables*)

1.  $H_{herd}(i)$ :  
*herding of households on stock  $i$  at day  $t$ .*
2.  $I_{herd}(i)$ :  
*herding of individuals on stock  $i$  at day  $t$ .*
3.  $F_{herd}(i)$ :  
*herding of foreigners on stock  $i$  at day  $t$ .*

# CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

Panel A: Full sample			
	Individuals	Institutions	Foreigners
Mean	0.1143 (0.0014)	0.1879 (0.0022)	0.1743 (0.0017)
Median	0.1132	0.1815	0.1759
SD	0.0274	0.0456	0.0365
Min	0.0241	-0.0413	0.0085
Max	0.1889	0.4675	0.3709

Table 11A

Panel B: Large stocks			
	Individuals	Institutions	Foreigners
Mean	0.1268 (0.0017)	0.1667 (0.0019)	0.1528 (0.0020)
Median	0.1269	0.1619	0.1507
SD	0.0243	0.0278	0.0309
Min	0.0222	0.1149	0.0095
Max	0.1875	0.2649	0.2554

Table 11B

# CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

Panel C: Small stocks			
	Individuals	Institutions	Foreigners
Mean	0.0949 (0.0014)	0.2123 (0.0035)	0.1987 (0.0020)
Median	0.0943	0.2107	0.1968
SD	0.0205	0.0509	0.0288
Min	0.0402	-0.0403	0.1252
Max	0.1556	0.4678	0.3722

Table 11C

## 4.4 Other predictors

### 4.4.1 Intraday volatility

Among several estimators designed to gauge intraday volatility, we employ the most popular estimator, the realized variance as introduced in Andersen, Bollerslev, Diebold, and Ebens (2001) [2]. To calculate the realized variance for stock  $i$  on day  $t$ , we first calculate five-minute returns by taking the log differences of prices observed at the end of each five-minute interval utilizing our transaction tick data. The realized variance of stock  $i$  on day  $t$ , then, is defined as the sum of the squared five-minute returns.

**Predictors 3.** (*Volatility variable*)

1.  $R.V(i)$ :  
a realized variance of a stock  $i$  at day  $t$ .

### 4.4.2 Predictors related to returns

In this section, we introduce predictors related to various notion of return. To this end, we first the notion of Volume Weighted Average Price (VWAP).

**Definition 4.4.1.**

$$VWAP_{i,t} = \frac{\sum_i Q_{i,t} P_{i,t}}{\sum_i Q_{i,t}}, \quad (4.4.1)$$

where  $Q_{i,t}$  is the quantity that is traded by traders of at price  $P_{i,t}$  at day  $t$ .

**Predictors 4.** (*Return variables*)

1.  $R(i)$ :  
the return of a stock  $i$  calibrated from closed price at day  $t - 1$  to closed price at day  $t$ .
2.  $VWAP2Close(i)$ :  
the return of a stock  $i$  calibrated from VWAP to closed price at day  $t$ .

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

3.  $VWAP(Hbuy)2Close$ :  
*the return of a stock  $i$  calibrated from VWAP payed by households for purchasing the stock  $i$  to closed price at day  $t$ .*
4.  $VWAP(Ibuy)2Close$ :  
*the return of a stock  $i$  calibrated from VWAP payed by institutional for purchasing the stock  $i$  to closed price at day  $t$ .*
5.  $VWAP(Fbuy)2Close$ :  
*the return of a stock  $i$  calibrated from VWAP payed by foreigners for purchasing the stock  $i$  to closed price at day  $t$ .*
6.  $VWAP(Hsell)2Close$ :  
*the return of a stock  $i$  calibrated from VWAP received by households for selling the stock  $i$  to closed price at day  $t$ .*
7.  $VWAP(Isell)2Close$ :  
*the return of a stock  $i$  calibrated from VWAP received by institutional for selling the stock  $i$  to closed price at day  $t$ .*
8.  $VWAP(Fsell)2Close$ :  
*the return of a stock  $i$  calibrated from VWAP received by foreigners for selling the stock  $i$  to closed price at day  $t$ .*

### 4.4.3 Predictors related to prices

In this section, we introduce predictors related to various notion of price. Every variable is calibrated in daily time horizon.

#### Predictors 5.

1.  $VWAP2close(i)$ :  
*the ratio of the VWAP of a stock  $i$  with respect to the closed price of the stock  $i$  at day  $t$ .*

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

2.  $VWAP(Hbuy)2VWAP(i)$ :  
*the ratio of the VWAP payed by households for purchasing a stock  $i$  with respect to the VWAP of the stock  $i$  at day  $t$ .*
3.  $VWAP(Ibuy)2VWAP(i)$ :  
*the ratio of the VWAP payed by institutions for purchasing a stock  $i$  with respect to the VWAP of the stock  $i$  at day  $t$ .*
4.  $VWAP(Fbuy)2VWAP(i)$ :  
*the ratio of the VWAP payed by foreigners for purchasing a stock  $i$  with respect to the VWAP of the stock  $i$  at day  $t$ .*
5.  $VWAP(Hsell)2VWAP(i)$ :  
*the ratio of the VWAP received by households for selling a stock  $i$  with respect to the VWAP of the stock  $i$  at day  $t$ .*
6.  $VWAP(Isell)2VWAP(i)$ :  
*the ratio of the VWAP received by institutions for selling a stock  $i$  with respect to the VWAP of the stock  $i$  at day  $t$ .*
7.  $VWAP(Fsell)2VWAP(i)$ :  
*the ratio of the VWAP received by foreigners for selling a stock  $i$  with respect to the VWAP of the stock  $i$  at day  $t$ .*

### 4.5 predictor model

To achieve outstanding performance by applying the deep learning algorithm to high dimensional data sets, we need a very large data set so-called 'Big-Data'(according to Goodfellow, Bengio, and Courville (2016) [18], for a supervised deep learning algorithm to produce similar or better performance than human, a dataset must contain at least 10 million labeled examples). Therefore, in case where our data set is not so much large, it is better to use shallow learning than deep learning. Among the various shallow learning techniques, in many case of high dimension data set, *Random Forest* produces the better

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

performance than linear models such as *Ridge Regression*, *Lasso Regression*, *Elastic Net*, etc.

One of the reasons that *Random Forest* produces the better performance than linear models is originated from the characteristics of tree model. Tree based non-linear models partition input space into several rectangles and assign an out value for each rectangle. By this way, tree based models can produce non-linear models which also include many types of linear models as their special cases.

However, the biggest drawback of the tree-based models is that, since each edge of the rectangles used in partitioning input space is parallel to an axis of input space, they do not work well when the true model is a linear model with a gradient vector not perpendicular to any axis in input space. This situation is described in Figure 4.1. Figure 4.1 shows that, to approximate a true separated line (red line), tree model have to split many times whereas linear model approximate the true separated line at once.

Therefore, by applying tree-based model to residuals after first fitting (removing) any linear part of true model, we can perform better than using tree-based model alone. In this section, we first introduce the two shallow learning techniques, *Random Forest* and *Elastic Net*, which are the main building blocks of our predictor model. We then describe our new model.

### 4.5.1 Model description

#### 4.5.1.1 Random forest

In this section, we briefly introduce *random forest* technique. *Decision tree* which is the algorithm based for *random forest* consists of 4 steps.

1. Growing: Find an optimal spitting rule for each note and grow the tree.  
Stop growing if stopping rule is satisfied.

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

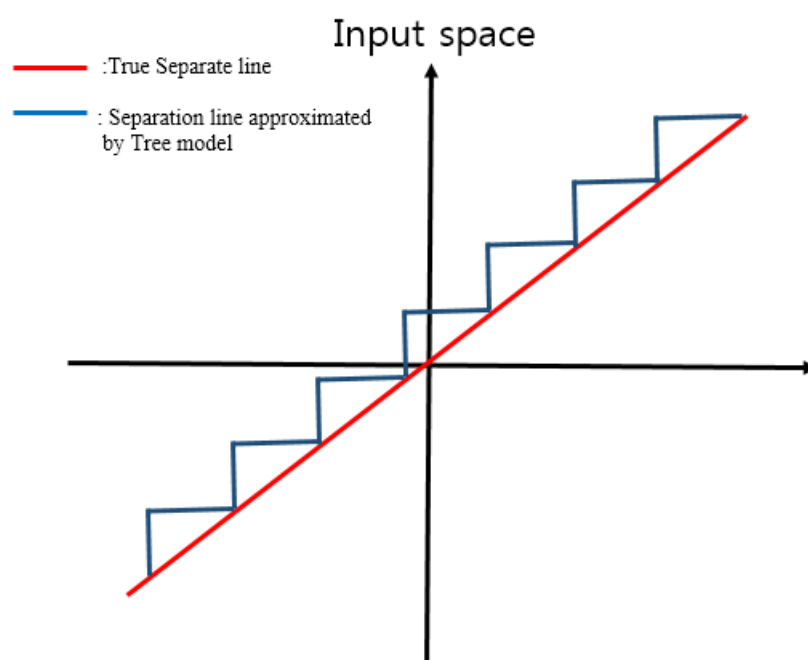


Figure 4.1



## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

2. Pruning: Remove nodes which increase prediction error or which have inappropriate inference rules. And also remove unnecessary (redundant) nodes.
3. Validation: Validation using gain chart, risk chart, test sample error, cross validation and etc. (to decide how much we prune the tree)
4. Interpretation and prediction: Interpret the constructed tree and predict

In *Decision Tree* model, to find an optimal splitting rules, uses an impurity measure. Given a node, impurity measure of a input variable is a measure of homogeneity of the target variable for the node. For example, in classification problem, a node in which the ratio of group 0 and 1 is 50:51 has a lower purity than a node in which the ratio of group 0 and 1 is 1:99.

**Splitting Rule of Tree Model.** *For each node, Decision Tree selects a split criterion which maximizes the sum of purities (minimize the sum of impurities) of the two child nodes.*

For a function  $\phi : [0, 1] \rightarrow [0, \infty)$  to be used impurity function, it should satisfy following conditions.

1.  $\phi(0) = \phi(1) = 0$ ,
2.  $\phi(1/2) = \text{maximum}$ ,
3.  $\phi(p) = \phi(1 - p)$ ,
4.  $\phi$  is concave.

For any impurity function,  $\phi : [0, 1] \rightarrow [0, \infty)$ , satisfying the above conditions, following property holds.

**Proposition 4.5.1.** *For given node  $t$ , let*

$$\Delta i(t) = \phi(p_t) - (\phi(p_{tR}) + \phi(p_{tL})),$$

*where  $\phi(p_t)$  is an impurity of the parent node,  $\phi(p_{tR})$  is an impurity of the right node, and  $\phi(p_{tL})$  is an impurity of the left node.*

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

*Proof.* See Proposition 4.4 in Breiman et al. (1984) [6]. □

Examples of impurity function are as follows.

- Classification model
  - $\chi^2$  statistics.
  - Gini index:  $\phi(p) = p(1 - p)$ .
  - Entropy index:  $\phi(p) = p \log p + (1 - p) \log(1 - p)$ .
- Regression model
  - F statistic of ANOVA.
  - Decrement of variance.

Based on impurity measures defined above, growing step consists of 2 steps.

1. Choose the optimal split for each node: for a given nodes, find splits minimizing the sum of impurities of child nodes. This maximizes the difference between impurity of parent node and sum of impurity of child nodes.
2. Choose the optimal node: find the split that not minimizes the sum of impurities of the child nodes, but maximizes difference of impurity between parent node and child nodes.

The most important feature of *random forest* is that the second step of the growing step in *decision tree* is replaced to random selection: in *random forest*, in growing step, an input variable is selected at random and a split position is calculated optimally. Then, take *bootstrap average* to acquire the final predictor model. one of the main advantage of *random forest* is that as Breiman (1999) [8] notes, it is more robust to output noise than other algorithms: Figure 4.2 , which is reproduced from Breiman (1999) [8], shows that Adaboost deteriorates substantially with 5% noise, while *random forest* generally show small changes.

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

<u>Data Set</u>	<u>Adaboost</u>	<u>Forest-RI</u>	<u>Forest-RC</u>
glass	1.6	.4	-.4
breast cancer	43.2	1.8	11.1
diabetes	6.8	1.7	2.8
sonar	15.1	-6.6	4.2
ionosphere	27.7	3.8	5.7
soybean	26.9	3.2	8.5
ecoli	7.5	7.9	7.8
votes	48.9	6.3	4.6
liver	10.3	-.2	4.8

Figure 4.2

### 4.5.1.2 Elastic Net

In this section, we briefly introduce Elastic Net. To this end, we first introduce *Ridge regression* and *Lasso regression*.

The *Ridge* estimator was proposed by Hoerl and Kennard (1970) [21] to resolve the problem of the least square estimator when  $p > n$ , where  $p$  : the number parameters,  $n$  : the number of observations.

**Definition 4.5.2.** (Ridge estimators)

$$\beta^{Ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{k=1}^p x_{ik} \beta_k \right)^2, \quad (4.5.1)$$

subject to

$$\sum_{k=1}^p \beta_k^2 \leq s,$$

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

or equivalently

$$\beta^{Ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{k=1}^p x_{ik} \beta_k \right)^2 + \lambda \sum_{k=1}^p \beta_k^2, \quad (4.5.2)$$

where  $\beta = (\beta_1, \dots, \beta_k)$ .

$s$ (or  $\lambda$ ) controls the complexity of the model. If  $s = 0$ , the model only includes the intercept term while the model becomes the full model when  $s = \infty$ . *Ridge* estimator is easily calibrated by Iterative Reweighted Least Square (IRLS). One of the disadvantages of *Ridge* estimator is that the interpretation of its' result is not easy since all predictor variables are used. As a estimator resolving this kinds of problem, *Lasso* estimator was proposed by Tibshirani (1996) [35].

**Definition 4.5.3.** (Lasso estimators)

$$\beta^{Lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{k=1}^p x_{ik} \beta_k \right)^2, \quad (4.5.3)$$

subject to

$$\sum_{k=1}^p |\beta_k| \leq s,$$

or equivalently

$$\beta^{Lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{k=1}^p x_{ik} \beta_k \right)^2 + \lambda \sum_{k=1}^p |\beta_k|, \quad (4.5.4)$$

where  $\beta = (\beta_1, \dots, \beta_k)$ .

The only difference between *Ridge* estimator and *Lasso* estimator is the penalty function. Whereas *Ridge* estimator uses  $L^2$  penalty function, *Lasso* estimator uses the  $L^1$  penalty. One of the main advantage of *Lasso* estimator over *Ridge* estimator is that it can do variable selection and shrinkage at the same time: the predictor model of *Lasso* estimator is sparse as we can see in

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

Figure 4.3, which is reproduced from Friedman, Hastie, Tibshirani (2001) [17]

Term	LS	Best Subset	Ridge	Lasso
Intercept	2.465	2.477	2.452	2.468
lcavol	0.680	0.740	0.420	0.533
lweight	0.263	0.316	0.238	0.169
age	-0.141		-0.046	
lbph	0.210		0.162	0.002
svi	0.305		0.227	0.094
lcp	-0.288		0.000	
gleason	-0.021		0.040	
pgg45	0.267		0.133	
Test Error	0.521	0.492	0.492	0.479
Std Error	0.179	0.143	0.165	0.164

Figure 4.3: coefficients of Least square estimator (LS), Best Subset selection estimator, Ridge estimator, Lasso estimator.

Therefore, *Lasso* estimator presents easier interpretation on its' result than *Ridge* estimator do. The variable selection by *Lasso* estimator can be intuitively explained by the Figure 4.4. Since the feasible set induced by the constraint of *Lasso* estimator has *corners*, the loss function of *Lasso* estimator is usually minimized at one of these *corners*. Hence, the coefficient corresponding to axis is shrunk to zero.

Since the property of sparse learning of *Lasso*, we have to optimize a non-differentiable objective function. There are at least three kinds of approaches to optimize *Lasso* estimator.

1. An approach based on the QP: calibration of *Lasso* estimator can be interpreted as a quadratic programming (QP) problem with linear constraints. This kinds of approach was first done by Tibshirani (1996) [35]. Later, Osborne (2000a, 2000b), ([29] [30]), Efron et al. (2004)[13] and Rosset and Zhu (2007) [32] developed the more efficient algorithms.

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

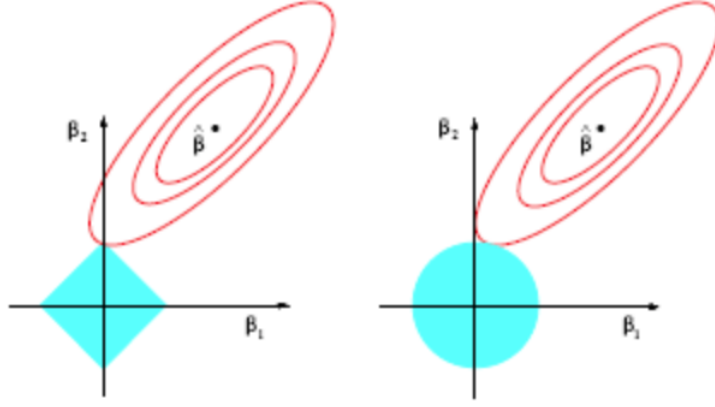


Figure 4.4: contours of the loss and constraint function of Lasso (left), Ridge (right) estimator.

2. An approach based on angle: LARS algorithm is presented by Efron et al. (2004) [13].
3. An approach based on gradient descent: solution path algorithm via sub-gradient is presented in Bühlmann, and van de Geer (2011) [8].

The statistical properties of *Lasso* estimator and *Ridge* estimator are as follows:

1. While *Ridge* estimator is not *persistent* (Kim (2005) [25]), *Lasso* estimator is *persistent* (Greenshtein and Ritov (2004) [20]) in the sense that

$$E(Y - \mathbf{X}'\hat{\beta})^2 - \operatorname{argmin}_{\beta \in R^{pn}} E(Y - \mathbf{X}'\beta)^2 \rightarrow 0$$

as  $p_n \rightarrow \infty$ .

2. *Lasso* estimator satisfies *the minimax optimality* in the sense

$$\sup_{\beta^*} E_{\beta^*}(\|\hat{\beta} - \beta^*\|) = O\left(\inf_{\hat{\gamma}} \sup_{\beta^*} E_{\beta^*}(\|\hat{\gamma} - \beta^*\|^2)\right),$$

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

where  $\hat{\gamma}$  runs over all estimators (Bickel et al. (2009) [3]).

3. The *selection consistency* hardly holds for *Lasso* estimator. In practically, *Lasso* estimator usually selects more variables than necessary variables (Zou (2006) [37]).

Sparse estimators such as *Lasso* produce good performance only when the true model we dealing with is sparse. However, when there are highly correlated predictor variables, in some case, the average of the predictor variables produces better performance than selection of a predictor as the following example 4.5.4 shows.

### Example 4.5.4.

- True model is given as follows.

$$Y = F + \varepsilon$$

where  $F \sim N(0, 1)$ ,  $\varepsilon \sim N(0, \sigma^2)$ , and  $F$  and  $\varepsilon$  are mutually independent.

- A data set:  $(Y, X_1, X_2)$  where  $X_j = F + \varepsilon_j$ ,  $\varepsilon_j \sim N(0, 1)$ , and  $F$  and  $\varepsilon_j$  are mutually independent for  $j = 1, 2$ .

Then

$$\underset{\beta_1, \beta_2}{\operatorname{argmin}} E(Y - \beta_1 X_1 - \beta_2 X_2)^2 = (1/2, 1/2). \quad (4.5.5)$$

Since highly correlated predictors are frequently occurred in high dimensional problems, we need an estimator which can manipulate the sparsity of the solution. The *Elastic Net* is a candidate for such kinds of estimators. The main idea of the *Elastic Net* is to combine the *Ridge* and *Lasso*.

### Definition 4.5.5. (Elastic Net estimators)

$$\beta^{Elastic \ Net} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{k=1}^p x_{ik} \beta_k \right)^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2^2. \quad (4.5.6)$$

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

Practically, it is well known that in most case, *Elastic Net* produces the best performance among the high dimensional linear shrinkage methods.

### 4.5.1.3 Our new model: two step learning (residual fitting)

We propose a regression model combining an advantage of high dimensional linear model (*Elastic Net*), and an advantage of tree based model (*Random Forest*), which is a high dimensional non linear model.

#### **Proposed Model: two-step learning.**

Given a  $N$  training data set  $\mathcal{D}_{train} = \{Y_i, X_i\}_{i=1}^N$ , and  $M$  test data set  $\mathcal{D}_{test} = \{Y_i, X_i\}_{i=1}^M$ .

- *Training step*

(Step1) *Target fitting: fitting the linear part of true model,  $\hat{f}_{linear}$ .*

*Based on the training set  $\mathcal{D}_{train}$ , estimate a linear predictor  $\hat{f}_{linear}$  via a high dimensional linear model (in our study, we use *Elastic Net*).*

(Step2) *Residual fitting: fitting the non linear part of true model,  $\hat{f}_{linear}$ .*

*Based on the residual set of the training set  $\mathcal{D}_{test}$ ,  $\mathcal{D}_{residual} = \{Y_i - \hat{f}_{linear}(X_i), X_i\}_{i=1}^N$ , estimate a non linear predictor  $\hat{f}_{nonlinear}$  via a high dimensional non linear model (in our study, we use *Random Forest*).*

- *Test step*

*For  $X_i \in \mathcal{D}_{test}$ , we predict  $Y_i$  as  $\hat{Y} = \hat{f}_{linear}(X_i) + \hat{f}_{nonlinear}(X_i)$ .*

### 4.5.2 Empirical Result

Our proprietary data sets can be divided into two kinds as follows.

- *Data set 1:*

*Data set 1* is composed of the intraday transaction data for all stock (2,131 stocks) listed on the KRX from February 1, 2008 through 30,



## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

December 2009 (479 trading days). The transaction data includes various information for every trade of our sample stocks during the sample period. For example, it contains the price and quantity of stocks sold or bought, trade time in milliseconds, trader type (which classifies the seller or the buyer into three groups: domestic individuals, domestic institutions, or foreign institutions.), and symbols that makes it possible to identify an **account** of seller and buyer for each trade uniquely.

- *Data set 2:*

*Data set 2* contains details of all transactions occurred from June 1, 2013 through May 31 2016 (736 trading days). The transaction data include several information for every trade of our sample stocks during the sample period. For example, it contains the price and quantity of stocks sold or bought, time of the transaction, a stock identifier, and trader type (which classifies the seller or the buyer into three groups: domestic individuals, domestic institutions, or foreign institutions.).

The biggest difference between *Data set 1* and *Data set 2* is that, whereas *Data set 1* contains the account information for each trade so that the information makes it possible for us to calibrate sell-buy ratio, *Data set 2* does not include the account information. However, all predictor variables introduced in this chapter, chapter 4, can be calculated for both of *Data set 1* and *Data set 2*. Therefore, this section consists of 2 parts. First, exploiting the *Data set 1*, we show that *our new model:two-step learning* described in the previous section has the best predictive power in forecasting sell-buy ratio. Second, after training *our new model:two-step learning* using *Data set 1*, we use the trained *our new model:2-step learning* in predicting stock return based on *Data set 2*, a data set does not have the account information.

### 4.5.2.1 SBR prediction

In this section, to show the superior performance of our new model (two-step learning) described in previous section, we compare performances of the three machine learning techniques, *Random Forest*, *Elastic Net*, and *our new model*:

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

*two-step learning*. Since only the *Data set 1* contains the account information of traders, we only use the *Data set 1* in predicting SBR (we apply the same kinds of filtrations in section 3.1 to the *Data set 1*). In predicting SBR, we do not use values of SBR itself but reassign values as following 2 steps.

1. for each day  $t$ , we partition all sample stocks into deciles based on SBR.
2. for each day  $t$ , for each stock  $i$ , we assign the number of decile the stock  $i$  belonging to the stock  $i$ 's SBR.

Table 12 reports the average prediction errors of the three machine learning techniques and their standard errors. The standard errors are calculated by *Bootstrap method*. The Table 12 shows that the average prediction error of *our new model: two-step learning* is the smallest of the three models and the result is statistically meaningful as the standard errors shows.

Prediction errors of the three machine learning techniques			
	Random Forest	Elastic Net	Our new model: 2-step learning
Average mean square errors	4.5115	4.5968	4.2857
Standard errors	0.0067	0.0098	0.0105

Table 12

### 4.5.2.2 Return prediction

The return prediction procedure consists of 2 steps as follows.

- Training step
  1. Apply the same kinds of filtrations in section 3.1 to the *Data set 1* as in the previous section.
  2. Fit the *our new model: two-step learning* to predict SBR using the *all Data set 1*.

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

- Test step
  1. Based on the *Data set 2*, for each day  $t$  of the sample period of *Data set 2*, predict SBR for all sample stocks at day  $t$  using the fitted model in training step.
  2. Based on the predicted SBR, for each day of sample period of *Data set 2*, construct a equal-weighted portfolio (*SBR portfolio*) of sample stocks with SBR larger than 9.

Figure 4.5 shows cumulative returns of *SBR Portfolio* and *Market Portfolio*. As the Figure 4.5 shows, *SBR Portfolio* shows the better performance than *Market Portfolio*, which means the *our new model: two-step learning* based on our stock price model has a predictive power in forecasting stock returns. Therefore, this result re convinces us that the our stock price model explain real market well.

## CHAPTER 4. RETURN PREDICTION VIA A MACHINE LEARNING TECHNIQUE

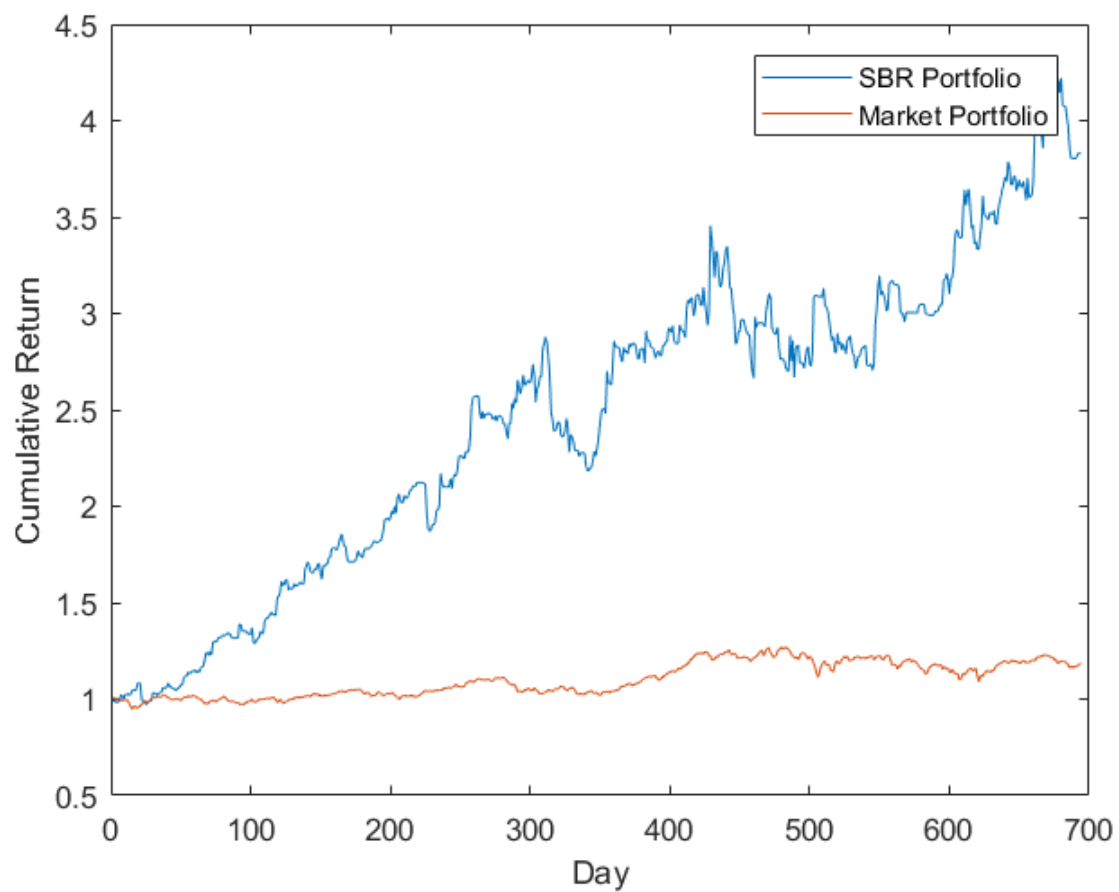


Figure 4.5: *SBR Portfolio vs Market Portfolio*

# Chapter 5

## Conclusion

Motivated by our prior studies (Chay and Kim (2017) [10]; Chay, Kim, and Lee (2017), [11]), which show the effect of investor dynamics on stock price processes, we suggest a new stock price process model in SDE form, which uses the SBR as the key variable. We then deduce closed-form solutions under the some linearity conditions. Our new model can be explained by information theory and empirical studies on traders' trading skills. Our new model presents two implications on *how the market work*.: First, in both the seller group and the buyer group, the stock price moves in the direction of the minority group, not the majority group, and the smaller the minority group, the larger the change in prices. Second, in both the seller group and the buyer group, traders follow (herd to) the behavior of the minority, and the smaller the minor group, the larger the herd.

We exploit our proprietary data set to show that the stock price process model we suggest explains the market well. The data set is composed of the intraday transaction data for all 2,131 stocks that were listed on the KRX from February 1, 2008, through 30, December 2009. The data contain symbols that make it possible to identify an account for each unique trade in the empirical test on our new stock price process model. The empirical result of the test shows that our model reflects the mechanism of the market well.

## CHAPTER 5. CONCLUSION

We use our model to predict stock prices via a *two-step machine-learning technique* (we invented) that combines a high-dimensional linear model (*Elastic Net*) and a high-dimensional non-linear model (*Random Forest*). We first show that the new machine-learning technique's predictive power is superior to machine-learning techniques that consist of one high-dimensional linear model or one high-dimensional non-linear model. Then we then show that we can predict stock returns by predicting the SBR using our new machine-learning technique, convincing us of the superiority of our stock price model in explaining market mechanisms.

# Bibliography

- [1] Admati, A. R., and P. Pfleiderer (1988) A theory of intraday patterns: Volume and price variability. *Review of Financial Studies*, 1, 3-40.
- [2] Andersen, T. G., T. Bollerslev, F. Diebold, and H. Ebens (2001) The distribution of stock return volatility. *Journal of Financial Economics*, 61, 43-76.
- [3] Bickel, P., Y. Ritov, and A. Tsybakov (2008). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37, 1705-1732.
- [4] Black, F. (1986) Noise *Journal of Finance*, 41, 529-543.
- [5] Bossaerts P., C. Frydman, and J. Ledyard (2014) The speed of information revelation and eventual price quality in markets with insiders: comparing two theories, *Review of Finance*, 18, 1-22.
- [6] Breiman, L., J.H. Friedman, R.A. Olshen, and C.I. Stone (1984) *Classification and regression trees*. Belmont, Calif.: Wadsworth.
- [7] Breiman, L. (1999) Random forests—random features, Technical Report 567, Department of Statistics, University of California, Berkley
- [8] Bühlmann, P., and S. van de Geer (2011) *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg.
- [9] Campbell (2006) Household finance. *Journal of Finance*, 61, 1553-1604.

## BIBLIOGRAPHY

- [10] Chay, J. B., and W. Kim (2017) Short-Term Trading Skills of Individuals, Institutions, and Foreigners: A New Approach Based on Relative Performance. Working Paper.
- [11] Chay, J. B., W. Kim, and Y. Lee. (2017) Intraday Volatility and Herding by Different Types of Investor: Evidence from an Emerging Market. Working Paper.
- [12] Clive Cookson (2013) "Time is money when it comes to microwaves". Financial Times.
- [13] Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. (2004) Least angle regression, *Annals of Statistics*, 32, 407-499.
- [14] Enke, D., S. Thawornwong (2005) The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, 29, 927-940.
- [15] Foster, F. D., and S. Viswanathan (1993) The effect of public information and competition on trading volume and price volatility. *The Review of Financial Studies*, 6, 23-56.
- [16] Foster, F. D., and S. Vishwanathan (1996) Strategic trading when agents forecast the forecasts of others. *Journal of Finance*, 51, 1437-1478.
- [17] Friedman, J.H., T. Hastie, R. Tibshirani (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, Heidelberg
- [18] Goodfellow, I., Y. Bengio, and A. Courville (2016) *Deep Learning* Cambridge MA: MIT Press
- [19] Greenshtein, E., Y. Ritov (2004) Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10, 971-988.



## BIBLIOGRAPHY

- [20] Grossman, S. J. and J. E. Stiglitz (1980) On the Impossibility of Informationally Efficient Markets. *The American Economic Review*, 70 (3), 393-408.
- [21] Hoerl, A. E. , R. W. Kennard (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems *Technometrics*, 12, 55-67.
- [22] Holden, C. W., and A. Subrahmanyam (1992) Long-lived private information and imperfect competition. *Journal of Finance*, 47, 247–270.
- [23] Holden, C. W., and A. Subrahmanyam (1994) Risk Aversion, Imperfect Competition, and Long-lived Information. *Economics Letters*, 44, 181-190.
- [24] Huang, W., Y. Nakamori, S.Y. Wang (2005) Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32, 2513-2522.
- [25] Kim, Y. (2005). Inconsistency of Ridge estimator in high dimensions. In <http://idea.snu.ac.kr>
- [26] Kyle, A.S. (1985) Continuous auctions and insider trading. *Econometrica*, 53, 1315-1335.
- [27] Lakonishok, J., A. Shleifer, and R. Vishny. (1992) The impact of institutional trading on stock prices, *Journal of Financial Economics*, 32, 23–43.
- [28] Nguyen,T.H., K. Shirai, J. Velcin (2015) Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42, 9603-9611.
- [29] Osborne, M.R., B. Presnell, and B.A. Turlach (2000a). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*, 20, 389-404.

## BIBLIOGRAPHY

- [30] Osborne, M.R., B. Presnell, and B.A. Turlach (2000b). On the LASSO and its dual. *Journal of the Computational and Graphical Statistics*, 9, 319-337.
- [31] Patel J.,S. Shah, P. Thakkar, K. Kotecha (2015) Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques. *Expert Systems with Applications*, 42, 259-268.
- [32] Rosset, S. and J. Zhu (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, 35.
- [33] Schnitzlein, Charles R. (1996) Price formation and market quality when the number and presence of insiders is unknown. *Review of Financial Studies*, 15, 1077–1110.
- [34] Stoffman, N. (2014) Who trades with whom? Individuals, institutions, and returns. *Journal of Financial Markets*, 21, 50–75.
- [35] Tibshirani, R. J. (1996) regression shrinkage and selection via the lasso. *journal of the royal statistical society Ser, B*, 58, 267-288.
- [36] Venezia, I., A. Nashikkar, and Z. Shapira. (2011) Firm specific and macro herding by professional and amateur investors and their effects on market volatility, *Journal of Banking & Finance*, 35 1599–1609
- [37] Zou, H. (2006) The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429.

## 국문초록

본 논문은 투자자들의 행동 역학을 반영한 새로운 주식 가격 과정 모형을 제시한다. 본 논문은 시장 작동 원리에 관한 두 가지 중요한 함의를 가진다. 첫째, 매도자 그룹과 매수자 그룹 중, 주식가격은 소수 그룹에 유리한 방향으로 움직이며, 소수 그룹의 규모가 작으면 작을수록 가격변화의 움직임은 커진다. 둘째, 매도자 그룹과 매수자 그룹 중, 투자자들은 소수그룹 포지션방향으로 움직이며, 소수 그룹의 규모가 작으면 작을 수록 그러한 허딩(herding)의 규모는 더 커진다. 또한 우리는 우리가 가지고 있는 고유한 데이터를 사용하여 본 논문이 제시한 모델이 실제 시장을 잘 설명함을 보인다. 마지막으로, 본 논문이 제시하는 모형을 바탕으로, 우리가 새롭게 개발한 기계학습 모형을 이용하여 주식가격을 예측할 수 있음을 보인다.

**주요어휘:** 주식 가격 과정, 내부자, 거래 기술, 기계학습, 주식가격 예측  
**학번:** 2013-30083